BAYESIAN METHODS FOR FUNCTIONAL AND TIME SERIES DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Daniel R. Kowal

August 2017

BAYESIAN METHODS FOR FUNCTIONAL AND TIME SERIES DATA

Daniel R. Kowal, Ph.D.

Cornell University 2017

We introduce new Bayesian methodology for modeling functional and time series data. While broadly applicable, the methodology focuses on the challenging cases in which (1) functional data exhibit additional dependence, such as time dependence or contemporaneous dependence; (2) functional or time series data demonstrate local features, such as jumps or rapidly-changing smoothness; and (3) a time series of functional data is observed sparsely or irregularly with non-negligible measurement error. A unifying characteristic of the proposed methods is the employment of the dynamic linear model (DLM) framework in new contexts to construct highly efficient Gibbs sampling algorithms.

To model dependent functional data, we extend DLMs for multivariate time series data to the functional data setting, and identify a smooth, time-invariant functional basis for the functional observations. The proposed model provides flexible modeling of complex dependence structures among the functional observations, such as time dependence, contemporaneous dependence, stochastic volatility, and covariates. We apply the model to multi-economy yield curve data and local field potential brain signals in rats.

For locally adaptive Bayesian time series and regression analysis, we propose a novel class of dynamic shrinkage processes. We extend a broad class of popular global-local shrinkage priors, such as the horseshoe prior, to the dynamic setting by allowing the local scale parameters to depend on the history of the shrinkage process. We prove that the resulting processes inherit desirable

shrinkage behavior from the non-dynamic analogs, but provide additional locally adaptive shrinkage properties. We demonstrate the substantial empirical gains from the proposed dynamic shrinkage processes using extensive simulations, a Bayesian trend filtering model for irregular curve-fitting of CPU usage data, and an adaptive time-varying parameter regression model, which we employ to study the dynamic relevance of the factors in the Fama-French asset pricing model.

Finally, we propose a hierarchical functional autoregressive (FAR) model with Gaussian process innovations for forecasting and inference of sparsely or irregularly sampled functional time series data. We prove finite-sample forecasting and interpolation optimality properties of the proposed model, which remain valid with the Gaussian assumption relaxed. We apply the proposed methods to produce highly competitive forecasts of daily U.S. nominal and real yield curves.

BIOGRAPHICAL SKETCH

Daniel Ryan Kowal was born in Albany, New York. After finishing high school at Salesianum School in Wilmington, Delaware, Daniel attended Washington University in St. Louis. While at Washington University, Daniel participated in the Pathfinder Program for Environmental Sustainability, completed a senior honors thesis *Applications of linear mixed effects models: an analysis of Missouri school data*, and graduated *summa cum laude* in mathematics with minors in computer science and legal studies. After graduating in 2012, Daniel entered the Cornell University Ph.D. program in statistics. During his graduate studies, Daniel co-authored publications in the *Journal of the American Statistical Association*, the *Journal of Business & Economic Statistics*, *Cellular and Molecular Bioengineering*, and the *Journal of Biomechanics*. He has received student paper awards from the *American Statistical Association* in both the *Section on Bayesian Statistical Science* and the *Nonparametric Statistics Section*. Following the completion of his Ph.D., Daniel will join the Rice University Department of Statistics as an assistant professor.

To my parents and my brother.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my co-advisers, Dr. David S. Matteson and Dr. David Ruppert, for their guidance, their time, and their commitment to my development as an independent researcher. I would also like to thank my undergraduate thesis advisor, Dr. Jimin Ding, for helping to set me on this path.

Second, I would like to thank my fellow Ph.D. students and friends, especially Dr. Amy Willis, Dr. David Sinclair, and Dr. William Nicholson. Both celebration and commiseration are unwritten prerequisites for graduate study, and their vital roles in each are greatly appreciated.

Third, I would like to thank my family, especially my parents, for persistently emphasizing the value of higher education and extracurricular learning, and my brother, for teaching me mathematics from an early age, despite my frequent protests.

And finally, I especially thank my wife, Dr. Marsha Kowal, for the encouraging notes, the travel packs, the early morning breakfast surprises, and most importantly, for her unwavering support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

We present Bayesian methodology for modeling functional and time series data. The methods are broadly applicable for (dependent) functional and time series data, but we focus in particular on the following challenging cases for which existing methods are inadequate:

1. Functional data with additional complex dependence, such as time dependence, contemporaneous dependence, stochastic volatility, covariates, and change points (Chapter 2);
2. Functional data, time series data, or regression functions with local features, such as jumps or rapidly-changing smoothness (Chapter 3); and
3. Forecasting and inference of functional time series data with sparsely or irregularly sampled curves and for curves sampled with non-negligible measurement error (Chapter 4).

A unifying characteristic of the proposed methods is the employment of the dynamic linear model (DLM) framework in new contexts to construct interpretable models and computationally efficient MCMC sampling algorithms. In particular, we develop highly efficient Gibbs sampling algorithms that build upon existing DLM sampling components for large blocks of parameters (e.g., Rue, 2001; Durbin and Koopman, 2002). The novel applications of DLMs include functional dynamic factor models, Bayesian trend filtering models, dynamic shrinkage processes (see Chapter 3), and functional autoregressive models. Importantly, the Bayesian framework permits joint estimation of the model

parameters and provides exact inference (up to MCMC error) on specific parameters.

The proposed methodology is motivated by important applications including multi-economy interest rate modeling, nominal and real yield curve forecasting, dynamic extensions of the Fama-French asset pricing model, irregular curve-fitting of CPU usage data, and local field potential brain signals in rats. The methods are evaluated through extensive simulations, and compared to state-of-the-art alternative estimators, with favorable results.

In Chapter 2, we present a Bayesian model for multivariate, dependent functional data, in which we extend DLMs for multivariate time series to the functional data setting. We also develop Bayesian spline theory in a more general constrained optimization framework. The proposed methods identify a time-invariant functional basis for the functional observations, which is smooth and interpretable. We apply the methodology to study the interactions of multi-economy yield curves during the recent global recession, and analyze local field potential brain signals in rats, for which we develop a multivariate functional time series approach for multivariate time-frequency analysis.

In Chapter 3, we propose a novel class of dynamic shrinkage processes for Bayesian time series and regression analysis. We extend a broad class of popular global-local shrinkage priors, such as the horseshoe prior, to the dynamic setting by allowing the local scale parameters to depend on the history of the shrinkage process. We prove that the resulting processes inherit desirable shrinkage behavior from the non-dynamic analogs, but provide additional locally adaptive shrinkage properties. The proposed dynamic shrinkage processes are widely applicable, particularly within the family of dynamic linear models. By express-

ing dynamic shrinkage processes on the log-scale, we adapt successful techniques from stochastic volatility modeling, and propose a Pólya-Gamma scale mixture representation to produce a highly efficient Gibbs sampling algorithm. We use the proposed processes to produce superior Bayesian trend filtering estimates and posterior credible intervals for irregular curve-fitting of minute-by-minute Twitter CPU usage data, and develop an adaptive time-varying parameter regression model to assess the efficacy of the Fama-French five-factor asset pricing model with momentum added as a sixth factor.

In Chapter 4, we develop a hierarchical Gaussian process model for forecasting and inference of functional time series data. Unlike existing methods, our approach is especially suited for sparsely or irregularly sampled curves and for curves sampled with non-negligible measurement error. The latent process is dynamically modeled as a functional autoregression (FAR) with Gaussian process innovations, with extensions for FAR(p) models with model averaging over the lag p. We propose a fully nonparametric dynamic functional factor model for the dynamic innovation process, with broader applicability and improved computational efficiency over standard Gaussian process models. We prove finite-sample forecasting and interpolation optimality properties of the proposed model, which remain valid with the Gaussian assumption relaxed. Extensive simulations demonstrate substantial improvements in forecasting performance and recovery of the autoregressive surface over competing methods, especially under sparse designs. We apply the proposed methods to forecast nominal and real yield curves using daily U.S. data. Real yields are observed more sparsely than nominal yields, yet the proposed methods are highly competitive in both settings.

CHAPTER 2

A BAYESIAN MULTIVARIATE FUNCTIONAL DYNAMIC LINEAR MODEL

Portions of this chapter were published in Kowal et al. (2016).

2.1 Introduction

We consider a multivariate time series of functional data. Functional data analysis (FDA) methods are widely applicable, including diverse fields such as economics and finance (e.g., Hays et al., 2012); brain imaging (e.g., Staicu et al., 2012); chemometric analysis, speech recognition, and electricity consumption (Ferraty and Vieu, 2006); and growth curves and environmental monitoring (Ramsay and Silverman, 2005). Methodology for independent and identically distributed (iid) functional data has been well-developed, but in the case of *dependent* functional data, the iid methods are not appropriate. Such dependence is common, and can arise via multiple responses, temporal and spatial effects, repeated measurements, missing covariates, or simply because of some natural grouping in the data (e.g., Horváth and Kokoszka, 2012). Here, we consider two distinct sources of dependence: time dependence for time-ordered functional observations and contemporaneous dependence for multivariate functional observations.

Suppose we observe multiple functions $Y_t^{(c)}(\tau)$, $c = 1, \ldots, C$, at time points $t = 1, \ldots, T$. Such observations have three dominant features:

(a) For each c and t, $Y_t^{(c)}(\tau)$ is a *function* of $\tau \in \mathcal{T}$;

(b) For each c and τ, $Y_t^{(c)}(\tau)$ is a *time series* for $t = 1, \ldots, T$; and

(c) For each t and τ, $Y_t^{(c)}(\tau)$ is a *multivariate* observation with outcomes $c = 1, \ldots, C$.

We assume that $\mathcal{T} \subseteq \mathbb{R}^d$ is compact, and focus on the case $d = 1$ in which τ is a scalar. However, our approach may be adapted to the more general setting.

We consider two diverse applications of multivariate functional time series (MFTS).

Multi-Economy Yield Curves: Let $Y_t^{(c)}(\tau)$ denote *multi-economy yield curves* observed on weeks $t = 1, \ldots, T$ for economies $c = 1, \ldots, C$, which refer to the Federal Reserve, the Bank of England, the European Central Bank, and the Bank of Canada. For a given currency and level of risk of a debt, the yield curve describes the interest rate as a function of the length of the borrowing period, or time to maturity, τ. Yield curves are important in a variety of economic and financial applications, such as evaluating economic and monetary conditions, pricing fixed-income securities, generating forward curves, computing inflation premiums, and monitoring business cycles (Bolder et al., 2004). We are particularly interested in the relationships among yield curves for the aforementioned globally-influential economies, and in how these relationships vary over time. However, existing FDA methods are inadequate to model the dynamic dependences among and between the yield curves for different economies, such as contemporaneous dependence, volatility clustering, covariates, and change points. Our approach resolves these inadequacies, and provides useful insights into the interactions among multi-economy yield curves (see Section 2.4.1).

Multivariate Time-Frequency Analysis: For multivariate time series, the peri-

odic behavior of the process is often the primary interest. *Time-frequency analysis* is used when this periodic behavior varies over time, which requires consideration of both the time and frequency domains (e.g., Shumway and Stoffer, 2000). Typical methods segment the multivariate time series into (overlapping) time bins within which the periodic behavior is approximately stationary; within each bin, standard frequency domain or spectral analysis is performed, which uses the multivariate discrete Fourier transform of the time series to identify dominant frequencies. Interestingly, although the raw signal in this setting is a multivariate time series, time-frequency analysis produces a MFTS: the multivariate discrete Fourier transform is a *function* of frequency τ for *time* bins $t = 1, \ldots, T$, where $c = 1, \ldots, C$ index the *multivariate* components of the spectrum. We analyze local field potential (LFP) data collected on rats, which measures the neural activity of local brain regions over time (Ljubojevic et al., 2013). Our interest is in the time-dependent periodic behavior of these local brain regions under different stimuli, and in particular the synchronization between brain regions. Our novel MFTS approach to time-frequency analysis provides the necessary multivariate structure and inference—which is unavailable in standard time-frequency analysis—to precisely characterize brain behavior under certain stimuli (see Section 2.4.2).

To model MFTS, we extend the hierarchical dynamic linear model (DLM) framework of Gamerman and Migon (1993) and West and Harrison (1997) for multivariate time series to the functional data setting. For smooth, flexible, and optimal function estimates, we extend Bayesian spline theory to a more general constrained optimization framework, which we apply for parameter identifiability. Our constraints are explicit in the posterior distribution via appropriate conditioning of the standard Bayesian spline posterior distribution, and the

corresponding posterior mean is the solution to an appropriate optimization problem. We implement an efficient Gibbs sampler to obtain samples from the joint posterior distribution, which provides exact (up to MCMC error) inference for any parameters of interest. The proposed hierarchical Bayesian *Multivariate Functional Dynamic Linear Model* has greater applicability and utility than related methods. It provides flexible modeling of complex dependence structures among the functional observations, such as time dependence, contemporaneous dependence, stochastic volatility, covariates, and change points, and can incorporate application-specific prior information.

The paper proceeds as follows. In Section 2.2, we present our model in its most general form. We develop our (factor loading) curve estimation technique in Section 2.3. In Section 2.4, we apply our model to the two applications discussed above and interpret the results. We also provide the details of our Gibbs sampling algorithm, present MCMC diagnostics for our applications, and include additional figures in Appendix A.

2.2 A Multivariate Functional Dynamic Linear Model

Suppose we observe functions $Y_t^{(c)}$: $\mathcal{T} \to \mathbb{R}$ at times $t = 1, \ldots, T$ for outcomes $c = 1, \ldots, C$, where $\mathcal{T} \subseteq \mathbb{R}$ is compact. We refer to the following model as the *Multivariate Functional Dynamic Linear Model* (MFDLM):

$$\begin{cases} \mathbf{Y}_t(\tau) = \mathbf{F}(\tau)\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t(\tau), & \left[\boldsymbol{\epsilon}_t(\tau)\middle|\mathbf{E}_t\right] \overset{indep}{\sim} N\left(\mathbf{0}, \mathbf{E}_t\right), \\ \boldsymbol{\beta}_t = \mathbf{X}_t\boldsymbol{\theta}_t + \boldsymbol{\nu}_t, & \left[\boldsymbol{\nu}_t\middle|\mathbf{V}_t\right] \overset{indep}{\sim} N(\mathbf{0}, \mathbf{V}_t), \\ \boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, & \left[\boldsymbol{\omega}_t\middle|\mathbf{W}_t\right] \overset{indep}{\sim} N(\mathbf{0}, \mathbf{W}_t), \end{cases} \tag{2.1}$$

where $\mathbf{Y}_t(\tau) = \left[Y_t^{(1)}(\tau), Y_t^{(2)}(\tau), \ldots, Y_t^{(C)}(\tau)\right]'$ is the C-dimensional vector of multivariate functional observations at time t evaluated at $\tau \in \mathcal{T}$; $\mathbf{F}(\tau)$ is the $C \times KC$ block matrix with $1 \times K$ diagonal blocks $\left[f_1^{(c)}(\tau), f_2^{(c)}(\tau), \ldots, f_K^{(c)}(\tau)\right]$ for $c = 1, \ldots, C$ of *factor loading curves* evaluated at $\tau \in \mathcal{T}$, with K the number of factors per outcome, and zeros elsewhere; $\boldsymbol{\beta}_t = \left[\beta_{1,t}^{(1)}, \ldots, \beta_{K,t}^{(1)}, \beta_{1,t}^{(2)}, \ldots, \beta_{K,t}^{(C)}\right]'$ is the KC-dimensional vector of *factors* that serve as the time-dependent weights on the factor loading curves; $\mathbf{X}_t$ is the known $KC \times p$ matrix of covariates at time t, where p is the total number of covariates; $\boldsymbol{\theta}_t$ is the p-dimensional vector of regression coefficients associated with $\mathbf{X}_t$; $\mathbf{G}_t$ is the $p \times p$ evolution matrix of the regression coefficients $\boldsymbol{\theta}_t$ at time t; and $\boldsymbol{\epsilon}_t(\tau)$, $\boldsymbol{\nu}_t$, and $\boldsymbol{\omega}_t$ are mutually independent error vectors with variance matrices $\mathbf{E}_t$, $\mathbf{V}_t$, and $\mathbf{W}_t$, respectively. We assume conditional independence of $[\boldsymbol{\epsilon}_t(\tau)|\mathbf{E}_t]$ over both $t = 1, \ldots, T$ and $\tau \in \mathcal{T}$; however, the latter assumption of independence over τ may be relaxed. We can immediately obtain a useful submodel of (2.1) by excluding covariates, $\mathbf{X}_t = \mathbf{I}_{CK \times CK}$, and removing a level of the hierarchy, $\mathbf{V}_t = \mathbf{0}_{CK \times CK}$, so that setting $\mathbf{G}_t = \mathbf{G}$ models $\boldsymbol{\beta}_t$ ($= \boldsymbol{\theta}_t$, almost surely) with a vector autoregression (VAR).

To understand (2.1), first note that the observation level of the model combines the *functional* component $\mathbf{F}(\tau)$ with the *multivariate time series* component $\boldsymbol{\beta}_t$. In scalar notation, we can write the observation level as

$$Y_t^{(c)}(\tau) = \sum_{k=1}^{K} f_k^{(c)}(\tau)\beta_{k,t}^{(c)} + \epsilon_t^{(c)}(\tau) \tag{2.2}$$

in which $\epsilon_t^{(c)}(\tau)$ are the elements of the vector $\boldsymbol{\epsilon}_t(\tau)$. In our construction, we can always write the observation level of (2.1) as (2.2); simplifications for the other levels will depend on the choice of submodel. For model identifiability, we require orthonormality of the factor loading curves:

$$\int_{\tau \in \mathcal{T}} f_k^{(c)}(\tau) f_j^{(c)}(\tau)\, d\tau = \mathbf{1}(k = j) \tag{2.3}$$

for $k, j = 1, \ldots, K$ and all outcomes $c = 1, \ldots, C$, where $\mathbf{1}(\cdot)$ is the indicator function. In addition, to ensure a unique and interpretable ordering of the factors $\beta_{1,t}^{(c)}, \ldots, \beta_{K,t}^{(c)}$ for each outcome $c = 1, \ldots, C$, we order the factor loading curves $f_1^{(c)}, \ldots, f_K^{(c)}$ by decreasing smoothness. We discuss our implementation of these constraints in Sections 2.3.2 and 2.3.3.

There are three primary interpretations of the model, which provide insight into useful extensions and submodels.

First, we can view (2.2) as a basis expansion of the functional observations $Y_t^{(c)}$, with a (multivariate) time series model for the basis coefficients $\beta_{k,t}^{(c)}$ to account for the additional dependence structures, such as common trends (see Section 2.4.1), stochastic volatility (see Section 2.4.1), and covariates. Since the identifiability constraint in (2.3) expresses orthonormality with respect to the L^2 inner product, we can interpret $\left\{ f_1^{(c)}, \ldots, f_K^{(c)} \right\}$ as an orthonormal basis for the functional observations $Y_t^{(c)}$. In contrast to common basis expansion procedures that assume the basis functions are known and only the coefficients need to be estimated (e.g., Bowsher and Meeks, 2008), we allow our basis functions $f_k^{(c)}$ to be estimated from the data. As a result, the $f_k^{(c)}$ will be more closely tailored to the data, which reduces the number of functions K needed to adequately fit the data. Conditional on the $f_k^{(c)}$, we can specify the $\boldsymbol{\beta}_t$- and $\boldsymbol{\theta}_t$-levels of (2.1) to appropriately model the remaining dependence among the $Y_t^{(c)}$. Using

this interpretation, we also note that (2.1) may be described as a multivariate dynamic (concurrent) functional linear model, and therefore extends a highly useful model in FDA (Cardot et al., 1999).

Similarly, we can interpret (2.1) as a dynamic factor analysis, which is a common approach in yield curve modeling (e.g., Hays et al., 2012; Jungbacker et al., 2013). Under this interpretation, the $\beta_{k,t}^{(c)}$ are dynamic *factors* and the $f_k^{(c)}$ are *factor loading curves* (FLCs); we will use this terminology for the remainder of the paper. Compared to a standard factor analysis, (2.1) has two major modifications: the factors $\beta_{k,t}^{(c)}$ are dynamic and therefore have an accompanying (multivariate) time series model, and the $f_k^{(c)}$ are functions rather than vectors.

Naturally, (2.1) has strong connections to a hierarchical DLM. Standard hierarchical DLM algorithms for sampling $\boldsymbol{\beta}_t$ and $\boldsymbol{\theta}_t$ assume that $\{\mathbf{F}, \mathbf{G}_t, \mathbf{X}_t, \mathbf{E}_t, \mathbf{V}_t, \mathbf{W}_t\}$ is known (e.g., Durbin and Koopman, 2002; Petris et al., 2009). Within our Gibbs sampler, we may *condition* on this set of parameters, and then use existing DLM algorithms to efficiently sample $\boldsymbol{\beta}_t$ and $\boldsymbol{\theta}_t$ with minimal implementation effort. Unconditionally, $\mathbf{F}$ is unknown, but we impose the necessary identifiability constraints; see Section 2.3 for more details. $\mathbf{G}_t$ may be known or unknown depending on the application, but in general it supplies the time series structure of the model (along with the time-dependent error variances): in Section 2.4.1, $\mathbf{G}_t = \mathbf{G}$ is unknown to allow for data-driven dependence among the multi-economy yield curves, and in Section 2.4.2, $\mathbf{G}_t = \mathbf{I}_{CK \times CK}$ is chosen to provide parsimonious time-domain smoothing. We assume that $\mathbf{X}_t$ is known, and may consist of covariates relevant to each outcome or can be chosen to provide additional shrinkage of $\boldsymbol{\beta}_t$ through $\boldsymbol{\theta}_t$. Although Gamerman and Migon (1993) suggest that $\dim(\boldsymbol{\theta}_t) < \dim(\boldsymbol{\beta}_t)$ for strict

dimension reduction in the hierarchy, we relax this assumption to allow for covariate information. Finally, we treat the error variance matrices as unknown, but typically there are simplifications available depending on the application and model choice. We discuss some examples in Section 2.4.

We must also specify a choice for K. In the yield curve application, two natural choices are $K = 3$ and $K = 4$ for comparison with the common parametric yield curve models: the Nelson-Siegel model (Nelson and Siegel, 1987) and the Svensson model (Svensson, 1994), both of which can be expressed as submodels of (2.1); see Diebold and Li (2006) and Laurini and Hotta (2010). More formally, we can treat K as a parameter and estimate it using reversible jump MCMC methods (Green, 1995), or select K using marginal likelihood. In particular, since we employ a Gibbs sampler, the marginal likelihood estimation procedure of Chib (1995) is convenient for many submodels of (2.1). For more complex models, DIC provides a less computationally intensive approach than either reversible jump MCMC or marginal likelihood, and is very simple to compute. In Appendix A, we discuss a fast procedure based on the singular value decomposition from our initialization algorithm which can be used to estimate a range of reasonable values for K.

2.3 Estimating the Factor Loading Curves

We would like to model the FLCs $f_k^{(c)}$ in a smooth, flexible, and computationally appealing manner. Clearly, the latter two attributes are important for broader applicability and larger data sets—including larger T, larger C, and larger $m_t^{(c)}$, where $m_t^{(c)}$ denotes the number of observation points for outcome c at time t.

The smoothness requirement is fundamental as well: as documented in Jungbacker et al. (2013), smoothness constraints can improve forecasting, despite the small biases imposed by such constraints. Smooth curves also tend to be more interpretable, since gradual trends are usually easier to explain than sharp changes or discontinuities.

However, there are some additional complications. First, we must incorporate the identifiability constraints, preferably without severely detracting from the smoothness and goodness-of-fit of the FLCs. We also have K curves to estimate for each outcome—or perhaps K curves common to all outcomes (see Section 2.3.4)—similar to the varying-coefficients model of Hastie and Tibshirani (1993), conditional on the factors $\beta_{k,t}^{(c)}$. Finally, the observation points for the functions $Y_t^{(c)}$ are likely different for each outcome c, and may also vary with time t.

2.3.1 Splines

A common approach in nonparametric and semiparametric regression is to express each unknown function $f_k^{(c)}$ as a linear combination of known basis functions, and then estimate the associated coefficients by maximizing a (penalized) likelihood (e.g., Wahba, 1990; Eubank, 1999; Ruppert et al., 2003). We use B-spline basis functions for their numerical properties and easy implementation, but our methods can accommodate other bases as well. For now, we ignore dependence on c for notational convenience; this also corresponds to either the univariate case ($C = 1$) or $C > 1$ with $\mathbf{E}_t$ diagonal and the FLCs assumed to be *a priori* independent for $c = 1, \ldots, C$ (see Section 2.3.4 for an important alterna-

tive). Following Wand and Ormerod (2008), we use cubic splines and the knot sequence $a = \kappa_1 = \ldots = \kappa_4 < \kappa_5 < \ldots < \kappa_{M+4} < \kappa_{M+5} = \ldots = \kappa_{M+8} = b$, with $\boldsymbol{\phi}_B = (\phi_1, \ldots, \phi_{M+4})$ the associated cubic B-spline basis, M the number of interior knots, and $\mathcal{T} = [a, b]$. While we could allow each f_k to have its own B-spline basis and accompanying sequence of knots, there is no obvious reason to do so. In our applications, we use $M = 20$ interior knots. For knot placement, we prefer a quantile-based approach such as the default method described in Ruppert et al. (2003), which is responsive to the location of observation points in the data yet is computationally inexpensive; however, equally-spaced knots may be preferable in some applications.

Explicitly, we write $f_k(\tau) = \boldsymbol{\phi}_B'(\tau)\boldsymbol{d}_k$, where $\boldsymbol{d}_k$ is the $(M+4)$-dimensional vector of unknown coefficients. Therefore, the function estimation problem is reduced to a vector estimation problem. In classical nonparametric regression, $\boldsymbol{d}_k$ is estimated by maximizing a penalized likelihood, or equivalently solving

$$\min_{\boldsymbol{d}_k} -2\log[\mathbf{Y}|\boldsymbol{d}_k] + \lambda_k \mathcal{P}(\boldsymbol{d}_k) \tag{2.4}$$

where $[\mathbf{Y}|\boldsymbol{d}_k]$ is a likelihood, $\mathcal{P}$ is a convex penalty function, and $\lambda_k \geq 0$. We express (2.4) as a log-likelihood multiplied by -2 so that for a Gaussian likelihood, (2.4) is simply a penalized least squares objective. For greater generality, we leave the likelihood unspecified, but later consider the likelihood of model (2.2). To penalize roughness, a standard choice for $\mathcal{P}$ is the L^2-norm of the second derivative of f_k, which can be written in terms of $\boldsymbol{d}_k$:

$$\mathcal{P}(\boldsymbol{d}_k) = \int_{\tau \in \mathcal{T}} \left[\ddot{f}_k(\tau)\right]^2 d\tau = \boldsymbol{d}_k' \boldsymbol{\Omega}_\phi \boldsymbol{d}_k \tag{2.5}$$

where $\ddot{f}_k$ denotes the second derivative of f_k and $\boldsymbol{\Omega}_\phi = \int_{\mathcal{T}} \ddot{\boldsymbol{\phi}}_B(\tau)\ddot{\boldsymbol{\phi}}_B{}'(\tau)\, d\tau$, which is easily computable for B-splines. With this choice of penalty, (2.4) balances goodness-of-fit with smoothness, where the trade-off is determined by

λ_k.

Since $\mathcal{P}$ is a quadratic in $\boldsymbol{d}_k$, for fixed λ_k, (2.4) is straightforward to solve for many likelihoods, in particular a Gaussian likelihood. Letting $\bar{\boldsymbol{d}}_k$ be this solution, we can estimate $f_k(\tau)$ for any $\tau \in \mathcal{T}$ with $\hat{f}_k(\tau) = \boldsymbol{\phi}'_B(\tau)\bar{\boldsymbol{d}}_k$. For a general knot sequence, the resulting estimator $\hat{f}_k$ is an O'Sullivan spline, or *O-spline*, introduced by O'Sullivan (1986) and explored in Wand and Ormerod (2008). In the special case of univariate nonparametric regression in which there is a knot at every observation point, $\hat{f}_k$ is a natural cubic smoothing spline (e.g., Green and Silverman, 1993). Alternatively, if we choose a sparser sequence of knots and set $\lambda_k = 0$, $\hat{f}_k$ is a regression spline (e.g., Ramsay and Silverman, 2005). O-splines are numerically stable, possess natural boundary properties, and can be computed efficiently (cf. Wand and Ormerod, 2008).

2.3.2 Bayesian Splines

Splines also have a convenient Bayesian interpretation (e.g., Wahba, 1978, 1983, 1990; Gu, 1992; Van der Linde, 1995; Berry et al., 2002). Returning to (2.4), we notably have a likelihood term and a penalty term, where the penalty is a function of only the vector of coefficients $\boldsymbol{d}_k$ and known quantities. Therefore, conditional on λ_k, the term $\lambda_k\mathcal{P}(\boldsymbol{d}_k)$ provides prior information about $\boldsymbol{d}_k$, for example that $f_k = \boldsymbol{\phi}'_B\boldsymbol{d}_k$ is smooth. Under this general interpretation, (2.4) combines the prior information with the likelihood to obtain an estimate of $\boldsymbol{d}_k$. A natural Bayesian approach is therefore to construct a prior for $\boldsymbol{d}_k$ based on the penalty $\mathcal{P}$, in particular so that the posterior mode of $\boldsymbol{d}_k$ is the solution to (2.4). For the most common settings in which the likelihood is Gaussian and the penalty $\mathcal{P}$

is (2.5), the posterior distribution of $\boldsymbol{d}_k$ will be Gaussian, so the posterior mean will also solve (2.4).

To construct a prior from $\mathcal{P}$, it is computationally and conceptually convenient to reparameterize $\boldsymbol{d}_k$ so that the penalty matrix $\boldsymbol{\Omega}_\phi$ is diagonal. Under a Gaussian prior, this corresponds to prior independence of the components of $\boldsymbol{d}_k$. The reparameterization will also affect the basis $\boldsymbol{\phi}_B$, but otherwise will leave the likelihood in (2.4) unchanged. Following Wand and Ormerod (2008), let $\boldsymbol{\Omega}_\phi = \mathbf{U}_\Omega \mathbf{D}_\Omega \mathbf{U}_\Omega'$ be the singular value decomposition of $\boldsymbol{\Omega}_\phi$, where $\mathbf{U}_\Omega' \mathbf{U}_\Omega = \mathbf{I}_{(M+4)\times(M+4)}$ and $\mathbf{D}_\Omega$ is a diagonal matrix with $M+2$ positive components. Denote the diagonal matrix of these positive entries by $\mathbf{D}_{\Omega,P}$ and let $\mathbf{U}_{\Omega,P}$ be the corresponding $(M+4)\times(M+2)$ submatrix of $\mathbf{U}_\Omega$. Using the reparameterized basis $\boldsymbol{\phi}'(\tau) = \left[1, \tau, \boldsymbol{\phi}_B'(\tau)\mathbf{U}_{\Omega,P}\mathbf{D}_{\Omega,P}^{-1/2}\right]$ and penalty $\boldsymbol{d}_k'\boldsymbol{\Omega}_D\boldsymbol{d}_k$ with $\boldsymbol{\Omega}_D = \text{diag}\,(0, 0, \lambda_k, \ldots, \lambda_k)$, the new solution $\hat{\boldsymbol{d}}_k$ to (2.4) satisfies $\hat{f}_k(\tau) = \boldsymbol{\phi}_B(\tau)\bar{\boldsymbol{d}}_k = \boldsymbol{\phi}'(\tau)\hat{\boldsymbol{d}}_k$; see Wand and Ormerod (2008) for more details. It is therefore natural to use the prior $\boldsymbol{d}_k \sim N(\mathbf{0}, \mathbf{D}_k)$, where $\mathbf{D}_k = \text{diag}\left(10^8, 10^8, \lambda_k^{-1}, \ldots, \lambda_k^{-1}\right)$ and $\lambda_k > 0$, which satisfies $\mathbf{D}_k^{-1} \approx \boldsymbol{\Omega}_D$. Notably, this prior is proper, yet is diffuse over the space of constant and linear functions—which are unpenalized by $\mathcal{P}$. This reparameterization is a common approach for fitting splines using mixed effects model software (e.g., Ruppert et al., 2003).

Since we assume conditional independence between levels of (2.1), our conditional likelihood for the FLCs is simply that of model (2.2), but we ignore dependence on c for now:

$$Y_t(\tau) = \sum_{k=1}^{K} \beta_{k,t} f_k(\tau) + \epsilon_t(\tau) = \sum_{k=1}^{K} \beta_{k,t} \boldsymbol{\phi}'(\tau)\boldsymbol{d}_k + \epsilon_t(\tau) \tag{2.6}$$

where $\epsilon_t(\tau) \overset{iid}{\sim} N(0, \sigma^2)$ for simplicity; the results are similar for more sophisticated error variance structures. In particular, (2.6) describes the distribution of

the functional data Y_t given the FLCs f_k (or $\boldsymbol{d}_k$), also conditional on $\beta_{k,t}$ and σ^2.

Under the likelihood of model (2.6) and the reparameterized (approximate) penalty $\boldsymbol{d}_k'\mathbf{D}_k^{-1}\boldsymbol{d}_k$, the solution to (2.4) conditional on $\boldsymbol{d}_j$, $j \neq k$ is given by $\hat{\boldsymbol{d}}_k = \mathbf{B}_k\mathbf{b}_k$ where $\mathbf{B}_k^{-1} = \mathbf{D}_k^{-1} + \sigma^{-2}\sum_{t=1}^{T}\beta_{k,t}^2\sum_{\tau\in\mathcal{T}_t}\boldsymbol{\phi}(\tau)\boldsymbol{\phi}'(\tau)$, $\mathbf{b}_k = \sigma^{-2}\sum_{t=1}^{T}\beta_{k,t}\sum_{\tau\in\mathcal{T}_t}\left[Y_t(\tau) - \sum_{j\neq k}\beta_{j,t}f_j(\tau)\right]\boldsymbol{\phi}(\tau)$, and $\mathcal{T}_t \subseteq \mathcal{T}$ denotes the discrete set of $|\mathcal{T}_t| = m_t$ observation points for Y_t at time t. Note that if $\mathcal{T}_t = \mathcal{T}_1$ for $t = 2, \ldots, T$, then $\mathbf{B}_k$ and $\mathbf{b}_k$ may be rewritten more conveniently in vector notation. Most importantly for our purposes, under the same likelihood induced by (2.6) and the prior $\boldsymbol{d}_k \sim N(\mathbf{0}, \mathbf{D}_k)$, the posterior distribution of $\boldsymbol{d}_k$ is multivariate Gaussian with mean $\hat{\boldsymbol{d}}_k$ and variance $\mathbf{B}_k$. For convenient computations, Wand and Ormerod (2008) provide an exact construction of $\boldsymbol{\Omega}_\phi$ and suggest efficient algorithms for $\hat{\boldsymbol{d}}_k$ based on the Cholesky decomposition; we provide more details in Appendix A.

To identify the ordering of the factors and FLCs in (2.2), we constrain the smoothing parameters $\lambda_1 > \lambda_2 > \cdots > \lambda_K > 0$. While other model constraints are available, this ordering constraint is particularly appealing: it sorts the FLCs f_k by decreasing smoothness, as characterized by the penalty function $\mathcal{P}$, and leads to a convenient prior distribution on the smoothing parameters λ_k. In the Bayesian setting, the smoothing parameters are equivalently the prior precisions of the penalized (nonlinear) components of $\boldsymbol{d}_k$. Letting $d_{k,j}$ denote the jth component of $\boldsymbol{d}_k$, the prior on the FLC basis coefficients is $d_{k,j} \overset{iid}{\sim} N(0, \lambda_k^{-1})$ for $j = 3, \ldots, M+4$. This is similar to the hierarchical setting of Gelman (2006), in which there are $M+2$ groups for each $\lambda_k, k = 1, \ldots, K$. Since $M+2$ is typically large, we follow the Gelman (2006) recommendation to place uniform priors on the group standard deviations $\lambda_k^{-1/2}, k = 1, \ldots, K$. Incorporating the ordering

constraint, the conditional priors are $\lambda_k^{-1/2} \sim \text{Uniform}\,(\ell_k, u_k)$, where $\ell_1 = 0$, $\ell_k = \lambda_{k-1}^{-1/2}$ for $k = 2, \ldots, K$, $u_k = \lambda_{k+1}^{-1/2}$ for $k = 1, \ldots, K-1$, and $u_K = 10^4$. The upper bound on $\lambda_K^{-1/2}$, and therefore all $\lambda_k^{-1/2}$, is chosen to equal the diffuse prior standard deviation of $d_{k,1}$ and $d_{k,2}$. The full conditional distributions of the smoothing parameters λ_k are Gamma $\left(\frac{1}{2}(M+1), \frac{1}{2}\sum_{j=3}^{M+4} d_{k,j}^2\right)$ truncated to (u_k^{-2}, ℓ_k^{-2}) for $k = 1, \ldots, K$, where we define $\ell_1^{-2} = \infty$. Notably, we avoid the diffuse Gamma prior on λ_k, which can be undesirably informative and is strongly discouraged by Gelman (2006). More generally, our approach provides a natural and data-driven method for estimating the smoothing parameters, yet does not inhibit inference. Details on the sampling of λ_k are provided in Appendix A.

2.3.3 Constrained Bayesian Splines

We extend the Bayesian spline approach to accommodate the necessary identifiability constraints for the MFDLM. For each $k = 1, \ldots, K$, we impose the orthonormality constraints $\int_{\mathcal{T}} f_k(\tau) f_j(\tau) = \mathbf{1}(k = j)$ for $j = 1, \ldots, K$. The unit-norm constraint preserves identifiability with respect to scaling, i.e., relative to the factors $\beta_{k,t}$ (up to changes in sign). The orthogonality constraints distinguish between pairs of FLCs, and in our approach identify the FLCs with distinct posterior distributions.

While other identifiability constraints are available for the f_k, orthonormality is appealing for a number of reasons. As discussed in Section 2.2, the orthonormality constraints suggest that we can interpret $\{f_1, \ldots, f_K\}$ as an orthonormal basis for the functional observations Y_t. As such, the orthogonality constraints

help eliminate any information overlap between FLCs, which keeps the total number of necessary FLCs to a minimum. Furthermore, the unit norm constraint allows for easier comparisons among the f_k. Of course, the f_k will be weighted by the factors $\beta_{k,t}$, so they can still have varying effects on the conditional mean of Y_t in (2.2). Finally, we can write the constraints conveniently in terms of the vectors $\boldsymbol{d}_k$ and $\boldsymbol{d}_j$:

$$\int_{\tau\in\mathcal{T}} f_k(\tau)f_j(\tau)\,d\tau = \int_{\tau\in\mathcal{T}} \boldsymbol{\phi}'(\tau)\boldsymbol{d}_k\boldsymbol{\phi}'(\tau)\boldsymbol{d}_j\,d\tau = \boldsymbol{d}_k'\mathbf{J}_\phi\boldsymbol{d}_j = \mathbf{1}(k=j) \tag{2.7}$$

for $j = 1,\ldots,K$, where $\mathbf{J}_\phi = \int_{\tau\in\mathcal{T}} \boldsymbol{\phi}(\tau)\boldsymbol{\phi}'(\tau)\,d\tau$ is easily computed for B-splines, and only needs to be computed once, prior to any MCMC sampling.

The addition of an orthogonality constraint to a (penalized) least squares problem has an intuitive regression-based interpretation, which we present in the following theorem:

Theorem 2.1. *Consider the penalized least squares objective* $\sigma^{-2}\sum_{i=1}^n(y_i - \mathbf{X}_i'\boldsymbol{d})^2 + \lambda\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d}$*, where* $y_i \in \mathbb{R}$*,* $\boldsymbol{d}$ *is an unknown* $(M+4)$*-dimensional vector,* $\mathbf{X}_i$ *is a known* $(M+4)$*-dimensional vector,* $\boldsymbol{\Omega}$ *is a known* $(M+4)\times(M+4)$ *positive-definite matrix, and* $\sigma^2, \lambda > 0$ *are known scalars. The solution is* $\hat{\boldsymbol{d}} = \mathbf{B}\mathbf{b}$*, where* $\mathbf{B}^{-1} = \lambda\boldsymbol{\Omega} + \sigma^{-2}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'$ *and* $\mathbf{b} = \sigma^{-2}\sum_{i=1}^n \mathbf{X}_i y_i$*. Now consider the same objective, but subject to the* J *linear constraints* $\boldsymbol{d}'\mathbf{L} = \mathbf{0}$ *for* $\mathbf{L}$ *a known* $(M+4)\times J$ *matrix of rank* J*. The solution is* $\tilde{\boldsymbol{d}} = \mathbf{B}\tilde{\mathbf{b}}$*, where* $\tilde{\mathbf{b}}$ *is the vector of residuals from the generalized least squares regression* $\mathbf{b} = \mathbf{L}\boldsymbol{\Lambda} + \boldsymbol{\delta}$ *with* $\mathbb{E}(\boldsymbol{\delta}) = 0$ *and* $\text{Var}(\boldsymbol{\delta}) = \mathbf{B}$*.*

Proof. The optimality of $\hat{\boldsymbol{d}}$ is a well-known result. For the constrained case, the Lagrangian is $\mathcal{L}(\boldsymbol{d},\boldsymbol{\Lambda}) = \sigma^{-2}\sum_{i=1}^n(y_i - \mathbf{X}_i'\boldsymbol{d})^2 + \lambda\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d} + \boldsymbol{d}'\mathbf{L}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the J-dimensional vector of Lagrange multipliers associated with the J linear constraints. It is straightforward to minimize $\mathcal{L}(\boldsymbol{d},\boldsymbol{\Lambda})$ with respect to $\boldsymbol{d}$ and obtain

the solution $\tilde{\boldsymbol{d}} = \mathbf{B}\tilde{\mathbf{b}} = \mathbf{B}(\mathbf{b} - \mathbf{L}\boldsymbol{\Lambda})$. Similarly, solving $\nabla\mathcal{L}(\tilde{\boldsymbol{d}}, \boldsymbol{\Lambda}) = \mathbf{0}$ for $\boldsymbol{\Lambda}$ implies that $\boldsymbol{\Lambda} = (\mathbf{L}'\mathbf{B}\mathbf{L})^{-1}\mathbf{L}'\mathbf{B}\mathbf{b}$, which is the solution to the generalized least squares regression of $\mathbf{b}$ on $\mathbf{L}$ with error variance $\mathbf{B}$. □

The result is interpretable: to incorporate linear constraints into a penalized least squares regression, we find $\tilde{\mathbf{b}}$ nearest to $\mathbf{b}$ under the inner product induced by $\mathbf{B}$ among vectors in the space orthogonal to $\text{Col}(\mathbf{L})$. In our setting, extending (2.4) under a Gaussian likelihood to accommodate the (linear) orthogonality constraints $\boldsymbol{d}_k'\mathbf{J}_\phi\boldsymbol{d}_j = 0$ for $j \neq k$ may be described via a regression of the unconstrained solution on the constraints. However, the unit norm constraint is nonlinear. This constraint affects the scaling but not the shape of f_k. Therefore, a reasonable approach is to construct a posterior distribution for $\boldsymbol{d}_k$ that respects the (linear) orthogonality constraints only, and then normalize the samples from this posterior to preserve identifiability. We provide more details in Appendix A.

To extend the unconstrained Bayesian splines of Section 2.3.2 to incorporate the orthogonality constraints, we write the constraints $\boldsymbol{d}_k'\mathbf{J}_\phi\boldsymbol{d}_j = 0$ for $j \neq k$ as the linear constraints in Theorem 2.1 with $\mathbf{L}_{[-k]} = (\mathbf{J}_\phi\boldsymbol{d}_1, \ldots, \mathbf{J}_\phi\boldsymbol{d}_{k-1}, \mathbf{J}_\phi\boldsymbol{d}_{k+1}, \ldots, \mathbf{J}_\phi\boldsymbol{d}_K)$ and $J = K - 1$. Using the full conditional posterior distribution $\boldsymbol{d}_k \sim N(\mathbf{B}_k\mathbf{b}_k, \mathbf{B}_k)$ from Section 2.3.2, we can additionally *condition* on the linear constraints $\boldsymbol{d}_k'\mathbf{L}_{[-k]} = \mathbf{0}$, and obtain the constrained full conditional distribution $\boldsymbol{d}_k \sim N(\tilde{\mathbf{B}}_k\mathbf{b}_k, \tilde{\mathbf{B}}_k)$, where $\tilde{\mathbf{B}}_k = \mathbf{B}_k - \mathbf{B}_k\mathbf{L}_{[-k]}(\mathbf{L}_{[-k]}'\mathbf{B}_k\mathbf{L}_{[-k]})^{-1}\mathbf{L}_{[-k]}'\mathbf{B}_k$. Conditioning on the orthogonality constraints is particularly interpretable in the Bayesian setting, and is convenient for posterior sampling; see Appendix A for more details. By comparison, Theorem 2.1 implies that the solution to (2.4) under the likelihood of model (2.6), the

penalty $\boldsymbol{d}_k' \mathbf{D}_k^{-1} \boldsymbol{d}_k$, and subject to the linear constraints $\boldsymbol{d}_k' \mathbf{L}_{[-k]} = \mathbf{0}$ is given by $\tilde{\boldsymbol{d}}_k = \mathbf{B}_k \tilde{\mathbf{b}}_k$, where $\tilde{\mathbf{b}}_k = \mathbf{b}_k - \mathbf{L}_{[-k]} \boldsymbol{\Lambda}_{[-k]}$ and $\boldsymbol{\Lambda}_{[-k]} = (\mathbf{L}_{[-k]}' \mathbf{B}_k \mathbf{L}_{[-k]})^{-1} \mathbf{L}_{[-k]}' \mathbf{B}_k \mathbf{b}_k$. Notably, $\tilde{\mathbf{B}}_k \mathbf{b}_k = \mathbf{B}_k \tilde{\mathbf{b}}_k = \tilde{\boldsymbol{d}}_k$, which is a useful result: by simply conditioning on the linear orthogonality constraints in the full conditional Gaussian distribution for $\boldsymbol{d}_k$, the posterior mean of the resulting Gaussian distribution solves the constrained regression problem of Theorem 2.1. In this sense, the identifiability constraints on f_k are enforced optimally.

2.3.4 Common Factor Loading Curves for Multivariate Modeling

Reintroducing dependence on c for the FLCs $f_k^{(c)}$, suppose that $C > 1$, so that our functional time series $Y_t^{(c)}$ is truly multivariate. If we wish to estimate *a priori* independent FLCs for each outcome c (with $\mathbf{E}_t$ diagonal), then we can sample from the relevant posterior distributions independently for $c = 1, \ldots, C$ using the methods of Section 2.3.3. The more interesting case is the *common factor loading curves model* given by $f_k^{(c)} = f_k$, so that all outcomes share a common set of FLCs. In the basis interpretation of the MFDLM, this corresponds to the assumption that the functional observations for all outcomes $Y_t^{(c)}$, $c = 1, \ldots, C$, $t = 1, \ldots, T$ share a common basis. We find this approach to be useful and intuitive, since it pools information across outcomes and suggests a more parsimonious model. Equally important, the common FLCs approach allows for direct comparison between factors $\beta_{k,t}^{(c)}$ and $\beta_{k,t}^{(c')}$ for outcomes c and c', since these factors serve as weights on the *same* FLC (or basis function) f_k. We use this model in both applications in Section 2.4.

The common FLCs model implies $f_k^{(c)}(\tau) = \boldsymbol{\phi}'_{(c)}(\tau)\boldsymbol{d}_k^{(c)} = f_k(\tau)$. However, since the FLCs for each outcome are identical, it is reasonable to assume that they have the same vector of basis functions $\boldsymbol{\phi}$, so $f_k^{(c)} = f_k$ is equivalent to $\boldsymbol{d}_k^{(c)} = \boldsymbol{d}_k$. Moreover, by writing $f_k^{(c)}(\tau) = \boldsymbol{\phi}'(\tau)\boldsymbol{d}_k$, we can use all of the observation points across all outcomes $c = 1, \ldots, C$ and times $t = 1, \ldots, T$, yet the parameter of interest, $\boldsymbol{d}_k$, will only be $(M + 4)$-dimensional.

Modifying our previous approach, we use the likelihood of model (2.2) with the simple error distribution $\epsilon_t^{(c)}(\tau) \overset{iid}{\sim} N(0, \sigma^2_{(c)})$. The implied full conditional posterior distribution for $\boldsymbol{d}_k$ is again $N(\tilde{\mathbf{B}}_k \mathbf{b}_k, \tilde{\mathbf{B}}_k)$, but now with $\mathbf{B}_k^{-1} = \mathbf{D}_k^{-1} + \sum_{c=1}^{C} \sigma^{-2}_{(c)} \sum_{t \in T^{(c)}} (\beta_{k,t}^{(c)})^2 \sum_{\tau \in \mathcal{T}_t^{(c)}} \boldsymbol{\phi}(\tau)\boldsymbol{\phi}'(\tau)$ and $\mathbf{b}_k = \sum_{c=1}^{C} \sigma^{-2}_{(c)} \sum_{t \in T^{(c)}} \beta_{k,t}^{(c)} \sum_{\tau \in \mathcal{T}_t^{(c)}} \left[Y_t^{(c)}(\tau) - \sum_{j \neq k} \beta_{j,t}^{(c)} f_j(\tau) \right] \boldsymbol{\phi}(\tau)$. For full generality, we allow the (discrete) set of times $T^{(c)}$ to vary for each outcome c and the (discrete) set of observation points $\mathcal{T}_t^{(c)}$ to vary with both time t and outcome c, with $|\mathcal{T}_t^{(c)}| = m_t^{(c)}$. Note that we reuse the same notation from Section 2.3.3 to emphasize the similarity of the multivariate results to the univariate (or *a priori* independent FLC) results. The common notation also allows for a more concise description of the sampling algorithm, which we present in Appendix A.

2.4 Data Analysis and Results

2.4.1 Multi-Economy Yield Curves

We jointly analyze *weekly* yield curves provided by the Federal Reserve (Fed), the Bank of England (BOE), the European Central Bank (ECB), and the Bank of Canada (BOC; Bolder et al. 2004) from late 2004 to early 2014 ($T = 490$ and $C =$

4). These data are publicly available and published on the respective central bank websites—and as such, we treat them as reliable estimates of the yield curves. For each outcome, the yield curves are estimated differently: the Fed uses quasi-cubic splines, the BOE uses cubic splines with variable smoothing parameters (Waggoner, 1997), the ECB uses Svensson curves, and the BOC uses exponential splines (Li et al., 2001). Therefore, the functional observations have already been smoothed, although by different procedures. The available set of maturities $\mathcal{T}_t^{(c)}$ is not the same across economies c, and occasionally varies with time t. The most frequent values of $m_t^{(c)}$, $t = 1, \ldots, T$, are 11 (Fed), 100 (BOE), 354 (ECB), and 120 (BOC), with maturities τ ranging from 1-3 months up to 300-360 months. To facilitate a simpler analysis, we let $Y_t^{(c)}(\tau)$ be the week-to-week *change* in the cth central bank yield curve on week t for maturity τ. Differencing the yield curves conveniently addresses the nonstationarity in the weekly data, and, because the yield curves are pre-smoothed, does not introduce any notable difficulties with time-varying observation points. We show an example of the multi-economy yield curves observed at adjacent times on July 29, 2011 and August 5, 2011, as well as the corresponding one-week change in Figure 2.1.

The literature on yield curve modeling is extensive. Yield curve models commonly adopt the Nelson-Siegel parameterization (Nelson and Siegel, 1987), often within a state space framework (e.g., Diebold and Li, 2006; Diebold et al., 2006, 2008; Koopman et al., 2010). Many Bayesian models also use the Nelson-Siegel or Svensson parameterizations (e.g., Laurini and Hotta, 2010; Cruz-Marcelo et al., 2011). However, the Nelson-Siegel parameterization does not extend to other applications, and often requires solving computationally intensive nonlinear optimization problems. Alternatively, Chib and Ergashev (2009) develop an arbitrage-free affine term structure model, which is similarly

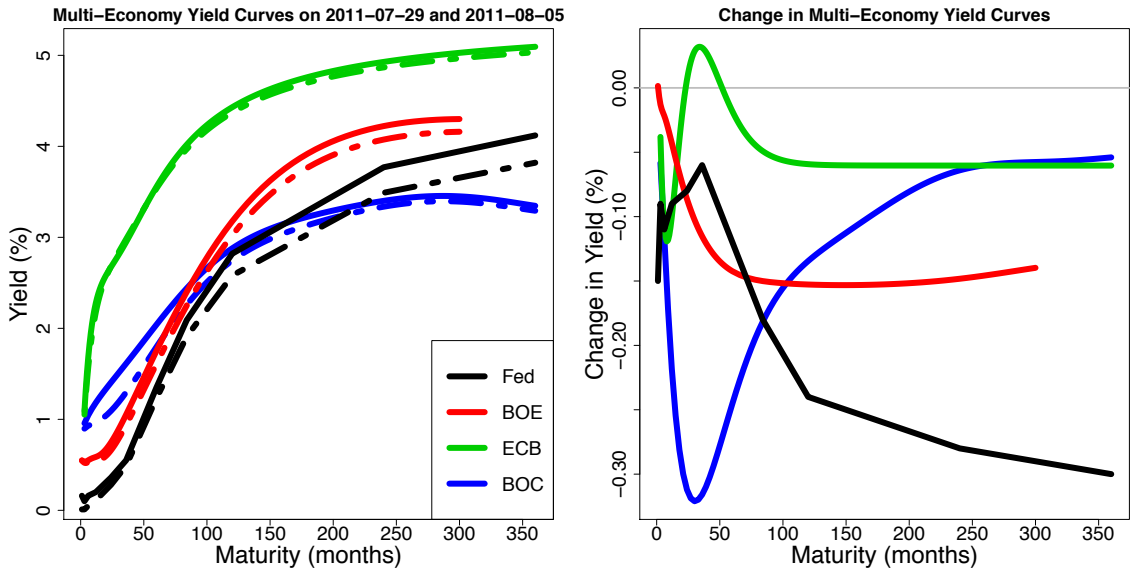


Figure 2.1: Multi-economy yield curves from July 29, 2011 (solid) and August 5, 2011 (dashed), together with the corresponding one-week change curves.

cast in a Bayesian state space framework. More similar to our approach are the Functional Dynamic Factor Model (FDFM) of Hays et al. (2012) and the Smooth Dynamic Factor Model (SDFM) of Jungbacker et al. (2013), both of which feature nonparametric functional components within a state space framework. The FDFM cleverly uses an EM algorithm to jointly estimate the functional and time series components of the model. However, the EM algorithm makes more sophisticated (multivariate) time series models more challenging to implement, and introduces some difficulties with generalized cross-validation (GCV) for estimation of the nonparametric smoothing parameters. The SDFM avoids GCV and instead relies on hypothesis tests to select the number and location of knots—and therefore determine the smoothness of the curves. However, this suggests that the smoothness of the curves depends on the significance levels used for the hypothesis tests, of which there can be a substantial number as $m_t^{(c)}$, C, or T grow large. By comparison, our smoothing parameters naturally depend on the data through the posterior distribution, which notably does *not*

create any difficulties for inference.

The multi-economy yield curves application is a natural setting for the common FLCs model of Section 2.3.4. First, since $f_k^{(c)} = f_k$ for $c = 1, \ldots, C$, the functional component of the MFDLM is the same for all economies, which helps reconcile the aforementioned different central bank yield curve estimation techniques. More specifically, the conditional expectations $\mu_t^{(c)}(\tau) \equiv \sum_{k=1}^{K} \beta_{k,t}^{(c)} f_k(\tau)$ are linear combinations of the *same* $\{f_1, \ldots, f_K\}$, and therefore are more directly comparable for $c = 1, \ldots, C$. Second, the common FLCs model is very useful when the set of observed maturities $\mathcal{T}_t^{(c)}$ varies with either outcome c or time t. Since the f_k are estimated using *all* of the observed maturities $\cup_{t,c} \mathcal{T}_t^{(c)}$, we notably do not need a missing data model for unobserved maturities at time t for economy c. In addition, for any $\tau \in \text{int range}\left(\cup_{t,c} \mathcal{T}_t^{(c)}\right)$, we may estimate $f_k(\tau)$ and $\mu_t^{(c)}(\tau)$ without any spline-related boundary problems—even when $\tau \notin \text{range}\left(\mathcal{T}_t^{(c)}\right)$. By comparison, non-common FLCs—or more generally, any linear combination of outcome-specific natural cubic splines—would impose a linear fit for $\tau \notin \text{range}\left(\mathcal{T}_t^{(c)}\right)$, which may not be reasonable for some applications.

The Common Trend Model

To investigate the similarities and relationships among the $C = 4$ economy yield curves, we implement the following parsimonious model for multivariate dependence among the factors:

$$\begin{cases} \beta_{k,t}^{(1)} = \omega_{k,t}^{(1)} \\ \beta_{k,t}^{(c)} = \gamma_k^{(c)} \beta_{k,t}^{(1)} + \omega_{k,t}^{(c)} \quad c = 2, \ldots, C \end{cases} \tag{2.8}$$

where $\gamma_k^{(c)} \in \mathbb{R}$ is the economy-specific slope term for each factor with the diffuse conjugate prior $\gamma_k^{(c)} \overset{iid}{\sim} N(0, 10^8)$. For the errors $\omega_{k,t}^{(c)}$, we use independent AR(r) models with time-dependent variances, which we discuss in more detail in Section 2.4.1. We also implement an interesting extension of (2.8) based on the autoregressive regime switching models of Albert and Chib (1993) and McCulloch and Tsay (1993) using the model $\beta_{k,t}^{(c)} = s_{k,t}^{(c)}(\gamma_k^{(c)}\beta_{k,t}^{(1)}) + \omega_{k,t}^{(c)}$, where $\left\{s_{k,t}^{(c)} : t = 1, \ldots, T\right\}$ is a discrete Markov chain with states $\{0, 1\}$. While this more complex model is not supported by DIC, it is a useful example of the flexibility of the MFDLM; we provide the details in Appendix A.

Letting $c = 1$ correspond to the Fed yield curve, we can use (2.8) to investigate how the factors $\beta_{k,t}^{(c)}$ for each economy $c > 1$ are *directly* related to those of the Fed, $\beta_{k,t}^{(1)}$. Since the U.S. economy is commonly regarded as a dominant presence in the global economy (e.g., Dées and Saint-Guilhem, 2011), the Fed yield curve is a natural and interesting reference point. Model (2.8) relates each economy $c > 1$ to the Fed using a regression framework, in which we regress $\beta_{k,t}^{(c)}$ on $\beta_{k,t}^{(1)}$ with AR(r) errors; since the yield curves were differenced, there is no need (or evidence) for an intercept. The slope parameters $\gamma_k^{(c)}$ measure the strength of this relationship for each factor k and economy c. In addition, we can investigate the residuals $\omega_{k,t}^{(c)}$ to determine times t for which $\beta_{k,t}^{(c)}$ deviated substantially from the linear dependence on $\beta_{k,t}^{(1)}$ assumed in model (2.8). Such periods of uncorrelatedness can offer insight into the interactions between the U.S. and other economies.

Stochastic Volatility Models

For the errors $\omega_{k,t}^{(c)}$ in (2.8), we use independent AR(r) models with time-dependent variances, i.e., $\omega_{k,t}^{(c)} = \sum_{i=1}^{r} \psi_{k,i}^{(c)} \omega_{k,t-i}^{(c)} + \sigma_{k,(c),t} z_{k,t}^{(c)}$ with $z_{k,t}^{(c)} \overset{iid}{\sim} N(0,1)$, $c = 1, \ldots, C$. The AR(r) specification accounts for the time dependence of the yield curves, while the $\sigma_{k,(c),t}^2$ model the observed volatility clustering. This latter component is important: in applications of financial time series, it is very common—and often necessary for proper inference—to include a model for the volatility (e.g., Taylor, 1994; Harvey et al., 1994). It is reasonable to suppose that applications of financial *functional* time series may also require volatility modeling; the weekly yield curve data provide one such example. Notably, our hierarchical Bayesian approach seamlessly incorporates volatility modeling, since, conditional on the volatilities, DLM algorithms require no additional adjustments for posterior sampling.

Within the Bayesian framework of the MFDLM, it is most natural to use a stochastic volatility model (e.g., Kim et al., 1998; Chib et al., 2002). Stochastic volatility models are parsimonious, which is important in hierarchical modeling, yet are highly competitive with more heavily parameterized GARCH models (Daníelsson, 1998). We model the log-volatility, $\log(\sigma_{(c),k,t}^2)$, as a stationary AR(1) process (for fixed c and k), using the priors and the efficient MCMC sampler of Kastner and Frühwirth-Schnatter (2014). We provide a plot of the volatilities $\sigma_{k,(c),t}^2$ and additional model details in Appendix A.

Results

We fit model (2.8) to the multi-economy yield curve data, using the the Kastner and Frühwirth-Schnatter (2014) model for the volatilities and setting $r = 1$, which adequately models the time dependence of the factors, with the diffuse stationarity prior $\psi_{k,1}^{(c)} \overset{iid}{\sim} N(0, 10^8)$ truncated to $(-1, 1)$. We use the common FLCs model of Section 2.3.4, and let $\mathbf{E}_t = \text{diag}\left(\sigma_{(1)}^2, \ldots, \sigma_{(C)}^2\right)$ with $\sigma_{(c)}^{-2} \overset{iid}{\sim} \text{Gamma}\,(0.001, 0.001)$. We prefer the choice $K = 4$, which corresponds to the number of curves in the Svensson model. However, since the observations $Y_t^{(c)}$ and the conditional expectations $\mu_t^{(c)}(\tau)$ are both smooth by construction, the errors $\epsilon_t^{(c)}$ are also smooth—and therefore correlated with respect to τ. To mitigate the effects of the error correlation, we increase the number of factors to $K = 6$, so that the fitted model (2.2) explains more than 99.5% of the variability in $Y_t^{(c)}(\tau)$. Since we are primarily interested in the first four factors, we fix $\gamma_k^{(c)} = 0$ for $k > 4$ in model (2.8), so the two additional factors for each outcome are modeled as independent AR(1) processes with stochastic volatility. We ran the MCMC sampler for $7,000$ iterations and discarded the first $2,000$ iterations as a burn-in. The MCMC sampler is efficient, especially for the factors $\beta_{k,t}^{(c)}$ and the common FLCs f_k; we provide the MCMC diagnostics in Appendix A.

In Figure 2.2, we plot the posterior means of the common FLCs f_k for $k = 1, \ldots, 4$. We can interpret these f_k as estimates of the time-invariant underlying functional structure of the yield curves shared by the Fed, the BOE, the ECB, and the BOC. The FLCs are very smooth, and the dominant hump-like features occur at different maturities—following from the orthonormality constraints—which allows the model to fit a variety of yield curve shapes. Interestingly, the estimated $f_1, f_2,$ and f_3 are similar to the level, slope, and curvature

functions of the Nelson-Siegel parameterization described by Diebold and Li (2006). Since the factors $\beta_{k,t}^{(c)}$ serve as weights on the FLCs f_k in (2.2), we may interpret the factors $\beta_{k,t}^{(c)}$—and therefore the slopes $\gamma_k^{(c)}$—based on these features of the yield curve explained by the corresponding f_k.

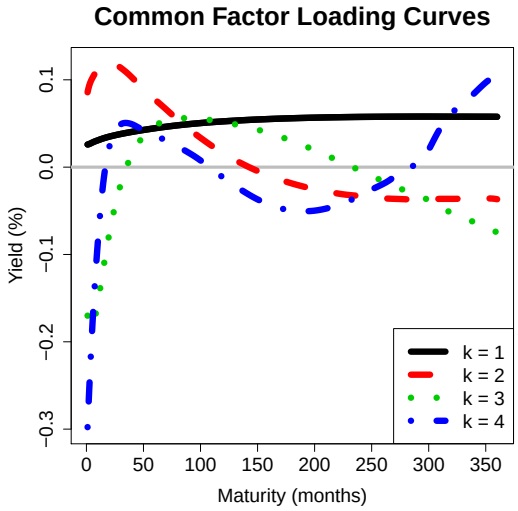


Figure 2.2: Posterior means of the common FLCs, $\{f_1, f_2, f_3, f_4\}$, as a function of maturity, τ.

In Table 1, we compute posterior means and 95% highest posterior density (HPD) intervals for $\gamma_k^{(c)}$, which measures the strength of the linear relationship between $\beta_{k,t}^{(c)}$ and $\beta_{k,t}^{(1)}$. For the level and slope factors $k = 1, 2$, the ECB is substantially less correlated with the Fed factors than are the BOE and BOC factors. For $k = 4$, the BOE, ECB, and BOC factors are nearly uncorrelated with the Fed factors.

Finally, we analyze the conditional standardized residuals from model (2.8), $r_{k,(c),t} = \left(\omega_{k,t}^{(c)} - \phi_{k,1}^{(c)}\omega_{k,t-1}^{(c)}\right)/\sigma_{k,(c),t} \overset{iid}{\sim} N(0,1)$, to determine periods of time t for which (2.8) is inadequate, which can indicate deviations from the assumed linear relationship between the Fed factors and the other economy factors. By computing the MCMC sample proportion of $r_{k,(c),t}^2 \sim \chi_1^2$ that exceed a critical

Economy	k = 1	k = 2	k = 3	k = 4
BOE	0.62 (0.57, 0.67)	0.72 (0.56, 0.89)	0.37 (0.27, 0.46)	0.03 (-0.03, 0.09)
ECB	0.39 (0.34, 0.45)	0.27 (0.11, 0.42)	0.44 (0.35, 0.52)	0.07 (0.00, 0.15)
BOC	0.61 (0.57, 0.65)	0.56 (0.47, 0.65)	0.49 (0.41, 0.58)	0.16 (0.08, 0.25)

Table 2.1: Posterior means and 95% HPD intervals for $\gamma_k^{(c)}$, which measures the strength of the linear relationship between $\beta_{k,t}^{(c)}$ and $\beta_{k,t}^{(1)}$.

value of the χ^2-distribution, e.g., the 95th percentile $\chi^2_{1,0.05} \approx 3.84$, we can obtain a simple estimate of the probability that $r^2_{k,(c),t}$ exceeds the critical value and, by that measure, is likely an outlier. We can compute a similar quantity for $\sum_{k=1}^4 r^2_{k,(c),t} \sim \chi^2_4$, which aggregates across factors $k = 1, \ldots, 4$. In Figure 2.3, we plot these MCMC sample proportions, restricted to the U.S. recession of December 2007 to June 2009. Around November 2008, there were outliers for all three economies for $k = 2, 3, 4$ and the aggregate, which suggests that the U.S. interest rate market may have behaved differently from the other economies during this time period. We are currently investigating an extension of model (2.8) to incorporate several important financial predictors as covariates, with a particular focus on the weeks during the recession.

2.4.2 Multivariate Time-Frequency Analysis for Local Field Potential

Local field potential (LFP) data were collected on rats to study the neural activity involved in *feature binding*, which describes how the brain amalgamates distinct sensory information into a single neural representation (Botly and De

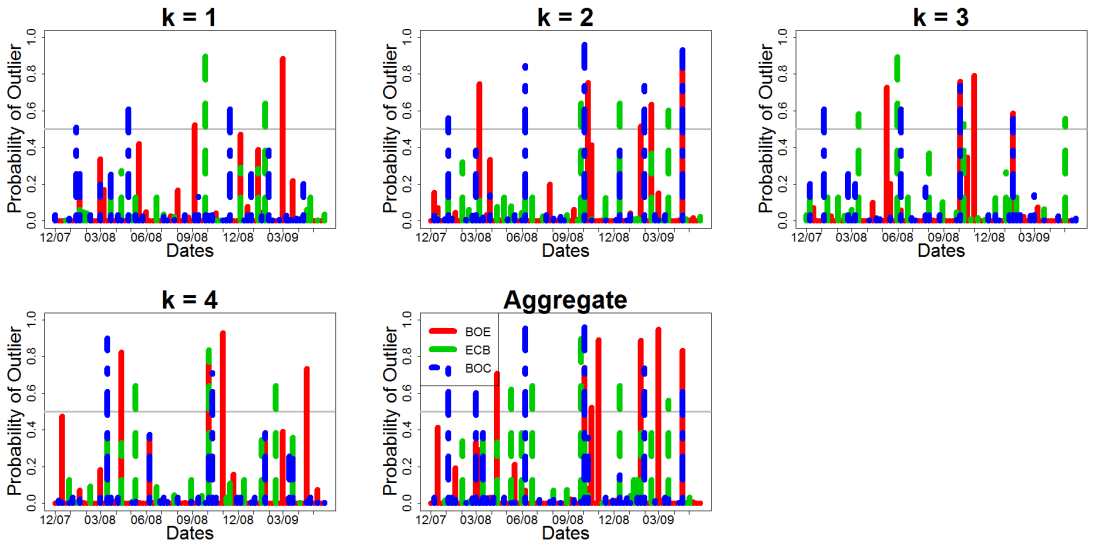
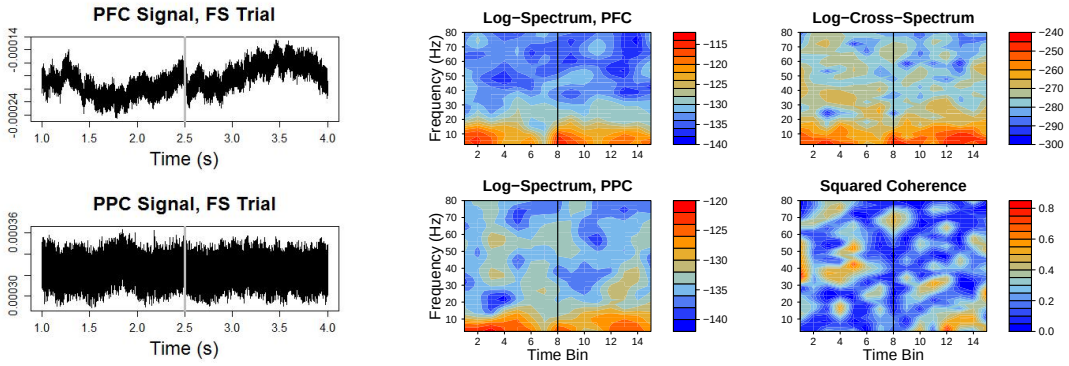


Figure 2.3: The MCMC sample proportions of $r^2_{k,(c),t}$ and $\sum_{k=1}^{4} r^2_{k,(c),t}$ that exceed the 95th percentile of the assumed χ^2-distributions.

Rosa, 2009; Ljubojevic et al., 2013). LFP uses pairs of electrodes implanted directly in local brain regions of interest to record the neural activity over time; in this case, the brain regions of interest are the prefrontal cortex (PFC) and the posterior parietal cortex (PPC). The rats were given two sets of tasks: one that required the rats to synthesize multiple stimuli in order to receive a reward (called *feature conjunction*, or FC), and one that only required the rats to process a single stimulus in order to receive a reward (called *feature singleton*, or FS). FC involves feature binding, while FS may serve as a baseline. The tasks were repeated in 20 trials each for FS and FC, during which electrodes implanted in the PFC and the PPC recorded the neural activity. Therefore, the raw data signal is a bivariate time series with 40 replications for each rat; we show an example of the bivariate signals for one such replication in Figure 2.4a. Each signal replicate is 3 seconds long, and has been centered around the behavior-based laboratory estimate of the time at which the rat processed the stimuli, which we denote by t^*.



(a) The bivariate LFP signal. (b) The associated (log-) spectra and squared coherence.

Figure 2.4: The raw LFP data from a rat during an FS trial. The vertical lines indicates the approximate time at which the rat processed the stimuli, t^*.

Our interest is in the time-dependent behavior of these bivariate signals and the interaction between them. A natural approach is to use *time-frequency analysis*; however, exact inference for standard time-frequency procedures is not available. An appealing alternative is to use time-frequency methods to transform the bivariate signal into a MFTS, which makes available the multivariate modeling and inference of the MFDLM.

Since the MFDLM provides smoothing in both the frequency domain $\mathcal{T}$ and the time domain T, we may use time-frequency preprocessing that provides minimal smoothing. For the time domain, we segment the signal into time bins of width one-eighth the length of the original signal, with a 50% overlap between neighboring bins to reduce undesirable boundary effects. Within each time bin, we compute the *periodograms* and *cross-periodogram* of the bivariate signal. Let $q_t^{(1)}(\tau)$ and $q_t^{(2)}(\tau)$ be the discrete Fourier transforms of the PFC and PPC signals, respectively, for time bin t evaluated at frequency τ, after removing linear trends. The periodograms are $I_t^{(c)}(\tau) = |q_t^{(c)}|^2$ for $c = 1, 2$ and the cross-periodogram is $I_t^{(3)}(\tau) = q_t^{(1)}\bar{q}_t^{(2)}$, where $\bar{q}_t^{(2)}$ is the complex conjugate of $q_t^{(2)}$. The cross-periodogram is generally complex-valued, and if the periodograms

are unsmoothed, then $|I_t^{(3)}(\tau)|^2 = I_t^{(1)}(\tau)I_t^{(2)}(\tau)$ is real-valued but clearly fails to provide new information (Bloomfield, 2004). This does not imply that the cross-periodogram is uninformative, but rather that some frequency domain smoothing of the periodograms is necessary.

Following Shumway and Stoffer (2000), we use a modified Daniell kernel to obtain the smoothed periodograms, or *spectra*. We subdivide each time bin into five segments, compute $I_t^{(c)}(\tau), c = 1, 2, 3$ within each segment, and then average the resulting periodograms using decreasing weights determined by the modified Daniell kernel. Denoting these spectra by $\tilde{I}_t^{(c)}(\tau)$, we let $Y_t^{(c)}(\tau) = \log\left(\tilde{I}_t^{(c)}(\tau)\right)$ for $c = 1, 2$, where the log-transformation is appealing because it is the variance-stabilizing transformation for the periodogram (Shumway and Stoffer, 2000). To account for the periodic dependence between signals, one choice is the log-cross-spectrum, $\log\left(|\tilde{I}_t^{(3)}(\tau)|^2\right)$. An appealing alternative is the *squared coherence* defined by $\kappa_t^2(\tau) \equiv |\tilde{I}_t^{(3)}(\tau)|^2/(\tilde{I}_t^{(1)}(\tau)\tilde{I}_t^{(2)}(\tau))$, which satisfies the constraints $0 \leq \kappa_t^2(\tau) \leq 1$ and is the frequency domain analog to the squared correlation (Bloomfield, 2004). Since (2.1) specifies that $Y_t^{(c)}(\tau) \in \mathbb{R}$, we transform the squared coherence and let $Y_t^{(3)}(\tau) = \Phi^{-1}(\kappa_t^2(\tau)) \in \mathbb{R}$, where $\Phi^{-1} : [0, 1] \to \mathbb{R}$ is a known monotone function; we use the Gaussian quantile function. We have found that fitting $Y_t^{(3)}(\tau)$ produces very similar results to fitting $\kappa_t^2(\tau)$ directly, yet in the transformed case, our estimate of the squared coherence $\Phi\left(\mu_t^{(3)}(\tau)\right)$ obeys the constraints. Because of our Bayesian approach, this transformation does not inhibit inference.

More generally, this procedure is applicable to ℓ-dimensional time series, which, including either the squared coherence or the cross-spectra, yields a $C = \ell(\ell + 1)/2$-dimensional MFTS. We show an example of the resulting MFTS from

a rat during an FS trial in Figure 2.4b. For completeness, we include the log-cross-spectrum, which is not a component of the MFTS.

MFDLM Specification

We use the common FLCs model of Section 2.3.4 accompanied by a random walk model for the factors:

$$\begin{cases} Y^{(c)}_{i,s,t}(\tau) = \sum_{k=1}^{K} \beta^{(c)}_{k,i,s,t} f_k(\tau) + \epsilon^{(c)}_{i,s,t}(\tau), & \left[\epsilon^{(c)}_{i,s,t}(\tau) \middle| \sigma^2_{(c)}\right] \overset{indep}{\sim} N(0, \sigma^2_{(c)}) \\ \boldsymbol{\beta}_{k,i,s,t} = \boldsymbol{\beta}_{k,i,s,t-1} + \boldsymbol{\omega}_{k,i,s,t}, & \left[\boldsymbol{\omega}_{k,i,s,t} \middle| \mathbf{W}_k\right] \overset{indep}{\sim} N(\mathbf{0}, \mathbf{W}_k) \end{cases} \tag{2.9}$$

where $\boldsymbol{\beta}_{k,i,s,t} = (\beta^{(1)}_{k,i,s,t}, \ldots, \beta^{(C)}_{k,i,s,t})'$, $Y^{(c)}_{i,s,t}$ are the log-spectra for $c = 1, 2$ and the probit-transformed squared coherences for $c = 3$, $i = 1, \ldots, 8$ index the rats, $s = 1, \ldots, 40$ index the trials for each rat, and $t = 1, \ldots, 15$ index the time bins for each trial. The joint indices (i, s, t) in (2.9) correspond to the time index t in (2.1), and are used to specify independence of the residuals $\boldsymbol{\omega}_{k,i,s,t}$ between rats and between trials. For each initial time bin $t = 1$, we let $\boldsymbol{\beta}_{k,i,s,1} \sim N(\mathbf{0}, 10^4 \mathbf{I}_{C\times C})$, since the corresponding observations are only time-ordered *within* a trial. The $C \times C$ factor covariance matrices $\mathbf{W}_k$ do not depend on the rat or the trial, and can help summarize the overall dependence among factors. For simplicity and parsimonious modeling, (2.9) assumes independence between $\boldsymbol{\omega}_{k,i,s,t}$ and $\boldsymbol{\omega}_{j,i,s,t}$ for $j \neq k \in \{1, \ldots, K\}$, but allows for correlation between outcomes for fixed k. The $\mathbf{W}_k$ control the amount of time domain smoothing for the factors and therefore for $\mu^{(c)}_{i,s,t}(\tau) \equiv \sum_{k=1}^{K} \beta^{(c)}_{k,i,s,t} f_k(\tau)$. For the error variances, we use the conjugate priors $\sigma^{-2}_{(c)} \overset{iid}{\sim} \text{Gamma}\,(0.001, 0.001)$ and $\mathbf{W}^{-1}_k \overset{iid}{\sim} \text{Wishart}((\rho R)^{-1}, \rho)$, with $R^{-1} = \mathbf{I}_{C\times C}$, the expected prior precision, and $\rho = C \geq \text{rank}(R^{-1})$. We provide the full conditional posterior distributions in Appendix A.

To determine the effects of feature binding, we compare the values of

$\mu_{i,s,t}^{(c)}(\tau)$ between the FS and FC trials. Letting $S_{i,FC}$ (respectively, $S_{i,FS}$) be the subset of FC (respectively, FS) trials for which rat i received the reward, we estimate posterior distributions for the sample means $\bar{\mu}_t^{(c)}(\tau) \equiv \frac{1}{8}\sum_{i=1}^{8}\left[\frac{1}{|S_{i,FC}|}\sum_{s\in S_{i,FC}}\mu_{i,s,t}^{(c)}(\tau) - \frac{1}{|S_{i,FS}|}\sum_{s'\in S_{i,FS}}\mu_{i,s',t}^{(c)}(\tau)\right]$ for $c = 1, 2$ and $\bar{\mu}_t^{(3)}(\tau) \equiv \frac{1}{8}\sum_{i=1}^{8}\left[\frac{1}{|S_{i,FC}|}\sum_{s\in S_{i,FC}}\Phi\left(\mu_{i,s,t}^{(3)}(\tau)\right) - \frac{1}{|S_{i,FS}|}\sum_{s'\in S_{i,FS}}\Phi\left(\mu_{i,s',t}^{(3)}(\tau)\right)\right]$. Therefore, we examine the difference in the log-spectra and the squared coherences between the FC and the FS trials, which we average over all rats and over all trials for which the rat responded *correctly* to the stimuli. This restriction is important, since it filters out unrepresentative trials, in particular FC trials for which feature binding may not have occurred.

Results

Since we observe functions in 15 time bins for 40 trials for 8 rats, the time-dimension of our 3-dimensional MFTS is $T = (15)(40)(8) = 4800$. We restrict the frequencies to $\mathcal{T} = [0.1, 80]$ Hz, which is the range of interest for this application and yields $m_t^{(c)} = 30$ for all c, t. Guided by DIC, we select $K = 10$. Alternatively, we could use a smaller value of K by increasing the initial smoothing of the log-spectra and the squared coherences, but would risk smoothing over important features. We ran the MCMC sampler for $7,000$ iterations and discarded the first $2,000$ iterations as a burn-in; see Appendix A for the MCMC diagnostics.

We compute 95% pointwise HPD intervals and posterior means for $\bar{\mu}_t^{(c)}(\tau)$, $c = 1, 2, 3$ and display the results as spectrogram plots; the plots for $c = 1, 2$ are in Appendix A, while $c = 3$ is in Figure 2.5. Regions of red or orange in the lower 95% HPD interval plots indicate a significant positive difference between the FC and FS trials, while regions of blue in the upper 95% HPD interval plots

indicate a significant negative difference. We are particularly interested in the time bins around t^*, which indicates the approximate time at which the stimuli were processed, and frequencies up to 40-50 Hz.

The averages of the differenced log-spectra, $\bar{\mu}_t^{(1)}(\tau)$ and $\bar{\mu}_t^{(2)}(\tau)$, describe how the distinct regions of the brain—the PFC and PPC, respectively—respond differently to stimuli that do or do not require feature binding. By comparison, the average of the differenced squared coherences, $\bar{\mu}_t^{(3)}(\tau)$, describes how these regions of the brain interact with each other under the different stimuli. Based on Figure 2.5, feature binding appears to be most strongly associated with greater squared coherence at frequencies in the Theta range (4-8 Hz), the Alpha range (8-13 Hz), and the Beta range (13-30 Hz) around t^*. This pattern persists in the power of both the PFC and PPC log-spectra plots, which suggests that these ranges of frequencies are important to the process of feature binding. Therefore, using the inference provided by the MFDLM, we conclude that during feature binding, the Theta, Alpha, and Beta ranges are associated with increased brain activity in both the PFC and the PPC, as well as greater synchronization between these regions.

2.5 Conclusions

The MFDLM provides a general framework to model complex dependence among functional observations. Because we separate out the functional component through appropriate conditioning and include the necessary identifiability constraints, we can model the remaining dependence using familiar scalar and multivariate methods. The hierarchical Bayesian approach allows us to incor-

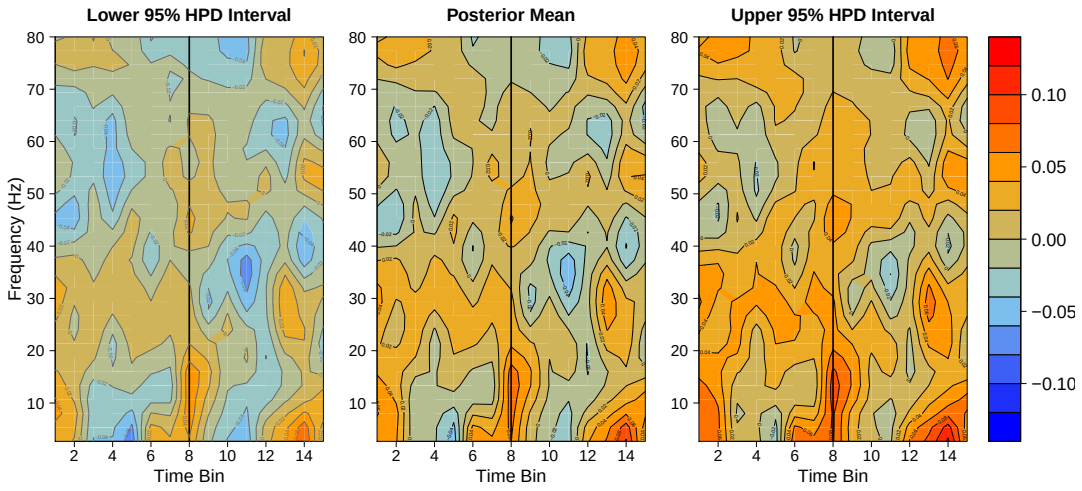


Figure 2.5: Pointwise 95% HPD intervals and the posterior mean for $\bar{\mu}_t^{(3)}$, which is the average difference in squared coherence between the FC and FS trials. The black vertical lines indicate the event time t^*.

porate interesting and useful submodels seamlessly, such as the common trend model of Section 2.4.1, the stochastic volatility model of Section 2.4.1, and the random walk model of Section 2.4.2. We combine Bayesian spline theory and convex optimization to model the functional component as a set of smooth and optimal curves subject to (identifiability) constraints. Using an efficient Gibbs sampler, we obtain posterior samples of all of the unknown parameters in (2.1), which allows us to perform inference on any parameters of interest, such as $\bar{\mu}_t^{(c)}$ in the LFP example.

Our two diverse applications demonstrate the flexibility and wide applicability of our model. The common trend model of Section 2.4.1 provides useful insights into the interactions among multi-economy yield curves, and our LFP example suggests a novel approach to time-frequency analysis via MFTS. In these applications, the MFDLM adequately models a variety of functional dependence structures, including time dependence, (time-varying) contempora-

neous dependence, and stochastic volatility, and may readily accommodate additional dependence structures, such as covariates, repeated measurements, and spatial dependence. We are currently developing an `R` package for our methods.

CHAPTER 3

DYNAMIC SHRINKAGE PROCESSES

3.1 Introduction

The global-local class of prior distributions is a popular and successful mechanism for providing shrinkage and regularization in a broad variety of models and applications. Global-local priors use continuous scale mixtures of Gaussian distributions to produce desirable shrinkage properties, such as (approximate) sparsity or smoothness, often leading to highly competitive and computationally tractable estimation procedures. For example, in the variable selection context, exact sparsity-inducing priors such as the spike-and-slab prior become intractable for even a moderate number of predictors. By comparison, global-local priors that shrink toward sparsity, such as the horseshoe prior (Carvalho et al., 2010), produce competitive estimators with greater scalability, and are validated by theoretical results, simulation studies, and a variety of applications (Carvalho et al., 2009; Datta and Ghosh, 2013; van der Pas et al., 2014). Unlike non-Bayesian counterparts such as the lasso (Tibshirani, 1996), shrinkage priors also provide adequate uncertainty quantification for parameters of interest (Kyung et al., 2010; van der Pas et al., 2014).

The class of global-local scale mixtures of Gaussian distributions (e.g., Car-

valho et al., 2010; Polson and Scott, 2010, 2012a) is defined as follows:

$$[\omega_t|\tau, \lambda_t] \overset{indep}{\sim} N(0, \tau^2\lambda_t^2), \quad t = 1, \ldots, T \tag{3.1a}$$

$$[\lambda_1^2, \ldots, \lambda_T^2] = \prod_{t=1}^{T} \left[\lambda_t^2 | \left\{\lambda_s^2\right\}_{s<t}\right] \tag{3.1b}$$

$$\sim \prod_{t=1}^{T} \pi(\lambda_t^2) \tag{3.1c}$$

where $\pi(\cdot)$ denotes a generic prior distribution and $\tau > 0$ is either endowed with its own prior distribution or estimated using empirical Bayes methods. Here (3.1c) follows from (3.1b) assuming the $\{\lambda_t\}$ are *a priori* independent and identically distributed (iid). The iid assumption is commonly made, but as we will argue below, it can be advantageous to forego the independence assumption. In what follows, only (3.1a) and (3.1b) will be assumed.

The prior in (3.1a)–(3.1c) is commonly paired with the likelihood $[y_t|\omega_t, \sigma^2] \overset{indep}{\sim} N(\omega_t, \sigma^2)$, but we will consider dynamic generalizations. In (3.1a), $\tau > 0$ controls the global shrinkage for all $\{\omega_t\}_{t=1}^T$, while λ_t tunes the local shrinkage for a particular ω_t. Such a model is particularly well-suited for sparse data: τ determines the global level of sparsity for $\{\omega_t\}_{t=1}^T$, while each λ_t allows for large absolute deviations of ω_t from its prior mean (zero). Careful choice of priors for λ_t^2 and τ^2 provide both robustness to large signals and adequate shrinkage of noise (e.g., Carvalho et al., 2010), so the framework of (3.1) is widely applicable.

In the dynamic setting, in which the observations y_t are time-ordered and t denotes a time index, it is natural to allow the local scale parameter, λ_t, to depend on the history of the shrinkage process $\{\lambda_s\}_{s<t}$. As a result, the probability of large (or small) deviations of ω_t from the prior mean (zero), as determined by λ_t, is informed by the previous shrinkage behavior $\{\lambda_s\}_{s<t}$. Such model-based

dependence may improve the ability of the model to adapt dynamically, which is important for time series estimation, forecasting, and inference. However, the standard global-local prior independence assumption in (3.1c) precludes dependence in the shrinkage process.

We propose to model the dynamic dependence of the process $\{\lambda_t\}$ in (3.1b) via a novel scale-mixture representation of stochastic volatility (SV) models. SV models for dynamic scale parameters are highly popular and successful, particularly in finance applications (e.g., Kim et al., 1998). In the standard SV model, $\{\lambda_t^2\}$ is modeled as an autoregressive process of order 1, or AR(1), on the log-scale. An important contribution of this manuscript is to extend the standard SV model to provide (1) direct extensions of popular shrinkage priors to the dynamic setting and (2) a highly efficient Gibbs sampling algorithm. We develop a log-scale representation of a broad class of global-local shrinkage priors, which provides a natural setting for modeling dynamic dependence. The proposed *dynamic shrinkage process* replaces the independence assumption (3.1c) with the dynamic evolution model

$$h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t, \quad \eta_t \overset{iid}{\sim} Z(\alpha, \beta, 0, 1) \tag{3.2}$$

where $h_t = \log(\tau^2\lambda_t^2)$, or equivalently $\tau^2 = \exp(\mu)$ and $\lambda_t^2 = \exp(h_t - \mu)$, and $Z(\alpha, \beta, \mu_z, \sigma_z)$ denotes the *Z-distribution* with density function

$$[z] = \big[\sigma B(\alpha,\beta)\big]^{-1}\big\{\exp\big[(z-\mu_z)/\sigma_z\big]\big\}^{\alpha}\big\{1+\exp\big[(z-\mu_z)/\sigma_z\big]\big\}^{-(\alpha+\beta)}, z \in \mathbb{R} \tag{3.3}$$

where $B(\cdot,\cdot)$ is the Beta function. When $\phi = 0$, model (3.2) reduces to the static setting, and implies an *inverted-Beta* prior for λ_t^2 (see Section 3.2.2 for more details). Notably, the class of priors represented in (3.2) includes the important shrinkage distributions in Table 3.1, in each case extended to the dynamic setting via an autoregression akin to the standard SV model.

$\alpha = \beta = 1/2$	Horseshoe Prior	Carvalho et al. (2010)
$\alpha = 1/2, \beta = 1$	Strawderman-Berger Prior	Strawderman (1971); Berger (1980)
$\alpha = 1, \beta = c-2, c > 0$	Normal-Exponential-Gamma Prior	Griffin and Brown (2005)
$\alpha = \beta \to 0$	(Improper) Normal-Jeffreys' Prior	Figueiredo (2003); Bae and Mallick (2004)

Table 3.1: Special cases of the inverted-Beta prior.

Despite the apparent complexity of the model, we develop a new Gibbs sampling algorithm that builds upon existing efficient sampling algorithms via a parameter expansion of model (3.2): a stochastic volatility sampler (Kim et al., 1998) and a Pólya-Gamma sampler (Polson et al., 2013). The resulting model is highly flexible, easy to implement, computationally efficient, and widely applicable.

For a motivating example, consider the minute-by-minute Twitter CPU usage data in Figure 3.1a (James et al., 2016). The data show an overall smooth trend interrupted by irregular jumps throughout the morning and early afternoon, with increased volatility from 16:00-18:00. It is important to identify both abrupt changes as well as slowly-varying intraday trends. To model these features, we combine the likelihood $y_t \overset{indep}{\sim} N(\beta_t, \sigma_t^2)$ with a standard SV model for the observation error variance, σ_t^2, and a *dynamic horseshoe process* as the prior on the second differences of the conditional mean, $\omega_t = \Delta^2\beta_t = \Delta\beta_t - \Delta\beta_{t-1}$, given by (3.2) with $\alpha = \beta = 1/2$ (see Section 3.3.2 for details). The dynamic horseshoe process either drives ω_t to zero, in which case β_t is locally linear, or leaves ω_t effectively unpenalized, in which case large changes in slope are permissible (see Figure 3.1b). The resulting posterior expectation of β_t and credible bands for the posterior predictive distribution of $\{y_t\}$ adapt to both irregular jumps and smooth trends (see Figure 3.1a).

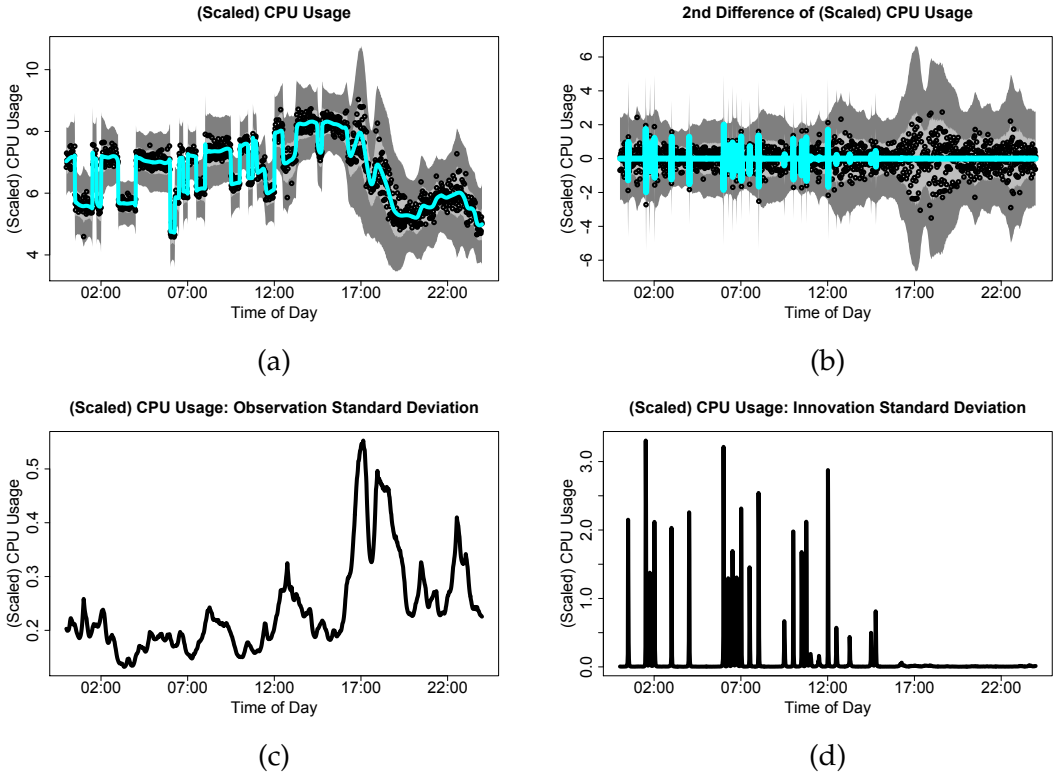


Figure 3.1: Bayesian trend filtering ($D = 2$) with dynamic horseshoe process innovations of minute-by-minute CPU usage data. (a) Observed data y_t (points), posterior expectation (cyan) of β_t, and 95% pointwise highest posterior density (HPD) credible intervals (light gray) and 95% simultaneous credible bands (dark gray) for the posterior predictive distribution of y_t. (b) Second difference of observed data $\Delta^2 y_t$ (points), posterior expectation of $\omega_t = \Delta^2 \beta_t$ (cyan), and 95% pointwise HPD intervals (light gray) and simultaneous credible bands (dark gray) for the posterior predictive distribution of $\Delta^2 y_t$. (c) Posterior expectation of time-dependent observation standard deviations, σ_t. (d) Posterior expectation of time-dependent innovation (prior) standard deviations, $\tau\lambda_t$.

For comparison, Figure 3.1 provides the posterior expectations of both the observation error standard deviations, σ_t (Figure 3.1c) and the prior standard deviations, $[\tau\lambda_t] = \exp(h_t/2)$ (Figure 3.1d). The horseshoe-like shrinkage behavior of λ_t is evident: values of λ_t are either near zero, corresponding to aggressive shrinkage of $\omega_t = \Delta^2\beta_t$ to zero, or large, corresponding to large absolute changes in the slope of β_t. Importantly, Figure 3.1 also provides motivation for a *dynamic* shrinkage process: there is clear *volatility clustering* of $\{\lambda_t\}$, in which the

shrinkage induced by λ_t persists for consecutive time points. The volatility clustering reflects—and motivates—the temporally adaptive shrinkage behavior of the dynamic shrinkage process.

Shrinkage priors and variable selection have been used successfully for time series modeling in a broad variety of settings. Belmonte et al. (2014) propose a Bayesian Lasso prior for shrinkage in dynamic linear models, while Korobilis (2013a) consider several (non-dynamic) scale mixture priors for time series regression. In both cases, the lack of a local (dynamic) scale parameter implies a time-invariant rate of shrinkage for each variable. Frühwirth-Schnatter and Wagner (2010) introduce indicator variables to discern between static and dynamic parameters, but the model cannot shrink adaptively for local time periods. Nakajima and West (2013) provide a procedure for local thresholding of dynamic coefficients, but the computational challenges of model implementation are significant. Chan et al. (2012) propose a class of time-varying dimension models, but due to the computational complexity of the model, only consider inclusion or exclusion of a variable for all times, which produces non-dynamic variable selection and a limited set of models.

Perhaps most comparable to the proposed methodology, Kalli and Griffin (2014) propose a class of priors which exhibit dynamic shrinkage using normal-gamma autoregressive processes. The Kalli and Griffin (2014) prior is a dynamic extension of the normal-gamma prior of Griffin and Brown (2010), and provides improvements in forecasting performance relative to non-dynamic shrinkage priors. However, the Kalli and Griffin (2014) model requires careful specification of several hyperparameters and hyperpriors, and the computation requires sophisticated adaptive MCMC techniques, which results in lengthy com-

putation times. By comparison, our proposed class of dynamic shrinkage processes is far more general, and includes the dynamic horseshoe process as a special case—which notably does not require tuning of sensitive hyperparameters. Furthermore, our proposed MCMC sampling algorithm combines existing samplers for large blocks of parameters, which produces a straightforward yet efficient Gibbs sampler, with computations linear in the number of time points.

We apply dynamic shrinkage processes to develop a dynamic fundamental factor model for asset pricing. We build upon the five-factor Fama-French model (Fama and French, 2015), which extends the three-factor Fama-French model (Fama and French, 1993) for modeling equity returns with common risk factors. We propose a dynamic extension which allows for time-varying factor loadings, possibly with localized or irregular features, and include a sixth factor, momentum (Carhart, 1997). Despite the popularity of the three-factor Fama-French model, there is not yet consensus regarding the necessity of all five factors in Fama and French (2015) or the momentum factor. Dynamic shrinkage processes provide a mechanism for addressing this question: within a time-varying parameter regression model, dynamic shrinkage processes provide the necessary flexibility to adapt to rapidly-changing features, while shrinking unnecessary factors to zero. Our dynamic analysis shows that with the exception of the market risk factor, no other risk factors are significant except for brief periods.

We introduce the dynamic shrinkage process in Section 3.2 and discuss relevant properties, including the Pólya-Gamma parameter expansion for efficient computations. In Section 3.3, we apply the prior to develop a more adaptive Bayesian trend filtering model for irregular curve-fitting, and we compare the

proposed procedure with competitive alternatives through simulations and a CPU usage application. We propose in Section 3.4 a time-varying parameter regression model with dynamic shrinkage processes for adaptive regularization and evaluate the model using simulations and an asset pricing example. In Section 3.5, we discuss the details of the Gibbs sampling algorithm, and conclude in Section 3.6. Proofs and additional details are in Appendix B.

3.2 Dynamic Shrinkage Processes

The proposed dynamic shrinkage process contains three prominent features: (1) a dynamic model for the local scale parameters, λ_t, via an autoregression on the log-scale; (2) a log-scale representation of a broad class of global-local priors to propagate desirable shrinkage properties to the dynamic setting; and (3) a Gaussian scale-mixture representation of the implied log-volatility evolution error to provide an efficient Gibbs sampling algorithm. In this section, we provide the relevant details regarding these features, and explore the properties of the resulting process.

3.2.1 Stochastic Volatility Models for Dynamic Scale Parameters

To extend the class of global-local scale mixtures of Gaussian distributions in (3.1) to the dynamic setting, we propose to model the local scale parameter, λ_t, using a *stochastic volatility* (SV) model (e.g., Kim et al., 1998). The SV model, which is the most common approach for modeling time-dependent random

scale parameters, introduces dynamic dependence via an AR(1) model for the log-variance (or log-volatility), as in model (3.2). Unlike model (3.2), standard SV models typically assume $\eta_t \overset{iid}{\sim} N(0, \sigma_\eta^2)$. A distinctive feature of the SV model is that it encourages volatility clustering, in which large (or small) variance—or shrinkage—persists for consecutive time points.

MCMC implementations of the SV model commonly represent the likelihood for h_t on the log-scale and approximate the resulting distribution using a known discrete mixture of Gaussian distributions (e.g., Kim et al., 1998). Importantly, the resulting approximation provides a framework for a fast and efficient MCMC sampler: conditional on the mixing component, the model for $\{h_t\}_{t=1}^T$ is a Gaussian dynamic linear model, and therefore $\{h_t\}_{t=1}^T$ may be sampled jointly in $\mathcal{O}(T)$ computations. We provide the relevant details in Section 3.5.

3.2.2 Log-Scale Representations of Global-Local Priors

Stochastic volatility models will not automatically exhibit desirable shrinkage behavior: we must consider appropriate distributions for μ and η_t. To illustrate this point, consider the standard SV model assumption for the evolution error distribution of log-volatility, $\eta_t \overset{iid}{\sim} N(0, \sigma_\eta^2)$. For the likelihood $y_t \sim N(\omega_t, 1)$ and the prior (3.1a), the posterior expectation of ω_t is $\mathbb{E}[\omega_t|\{y_s\}, \tau] = (1 - \mathbb{E}[\kappa_t|\{y_s\}, \tau])\, y_t$, where

$$\kappa_t \equiv \frac{1}{1 + \text{Var}\left[\omega_t|\tau, \lambda_t\right]} = \frac{1}{1 + \tau^2\lambda_t^2} \tag{3.4}$$

is the *shrinkage parameter*. As noted by Carvalho et al. (2010), $\mathbb{E}[\kappa_t|\{y_s\}, \tau]$ is interpretable as the amount of shrinkage toward zero *a posteriori*: $\kappa_t \approx 0$ yields minimal shrinkage (for signals), while $\kappa_t \approx 1$ yields maximal shrinkage to zero

(for noise). For the standard SV model and fixing $\phi = \mu = 0$ for simplicity, $\lambda_t = \exp(\eta_t/2)$ is log-normally distributed, and the shrinkage parameter has density $[\kappa_t] \propto \frac{1}{\kappa_t(1-\kappa_t)} \exp\left\{ -\frac{1}{2\sigma_\eta^2}\left[\log\left(\frac{1-\kappa_t}{\kappa_t}\right)\right]^2\right\}$. Notably, the density for κ_t approaches zero as $\kappa_t \to 0$ and as $\kappa_t \to 1$. As a result, direct application of the Gaussian SV model may overshrink true signals and undershrink noise.

By comparison, consider the horseshoe prior of Carvalho et al. (2010). The horseshoe prior is the special case of (3.1c) with $[\lambda_t] \overset{iid}{\sim} C^+(0,1)$, where C^+ denotes the half-Cauchy distribution. For fixed $\tau = 1$, the half-Cauchy prior on λ_t is equivalent to $\kappa_t \overset{iid}{\sim} \text{Beta}(1/2, 1/2)$, which induces a "horseshoe" shape for the shrinkage parameter (see Figure 3.2). The horseshoe-like behavior is ideal in sparse settings, since the prior density allocates most of its mass near zero (minimal shrinkage of signals) and one (maximal shrinkage of noise). Theoretical results, simulation studies, and a variety of applications confirm the effectiveness of the horseshoe prior (Carvalho et al., 2009, 2010; Datta and Ghosh, 2013; van der Pas et al., 2014).

To emulate the robustness and sparsity properties of the horseshoe and other shrinkage priors in the dynamic setting, we represent a general class of global-local shrinkage priors on the log-scale. As a motivating example, consider the special case of (3.1a) and (3.2) with $\phi = 0$: $\omega_t \overset{indep}{\sim} N(0, \tau^2\lambda_t^2)$ with $\log(\lambda_t^2) = \eta_t$. This example is illuminating: we equivalently express the (static) horseshoe prior by letting $\eta_t \overset{D}{=} \log \lambda_t^2$, where $\overset{D}{=}$ denotes equality in distribution. In particular,

$$\left[\lambda_t^2\right] \propto \left(\lambda_t^2\right)^{-1/2}\left(1+\lambda_t^2\right)^{-1}$$

implies

$$[\eta_t] = \pi^{-1}\exp(\eta_t/2)\left[1+\exp(\eta_t)\right]^{-1}$$

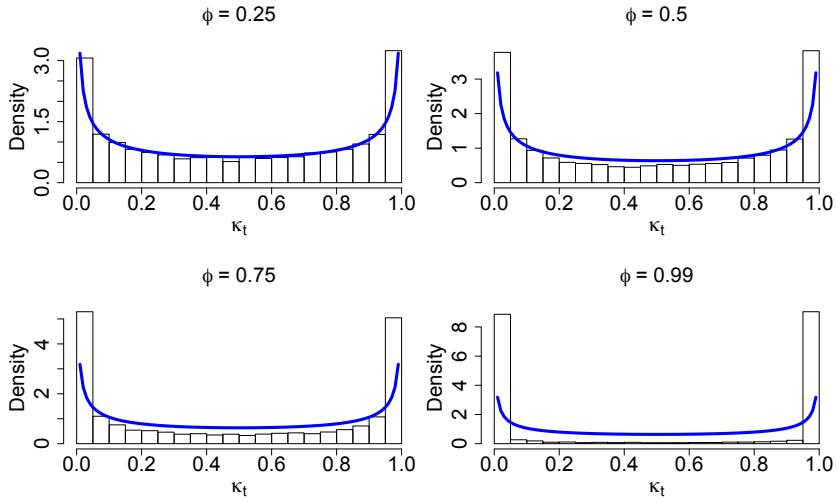


Figure 3.2: Simulation-based estimate of the stationary distribution of κ_t for various AR(1) coefficients ϕ. The blue line indicates the density of κ_t in the static ($\phi = 0$) horseshoe, $[\kappa] \sim \text{Beta}\,(1/2, 1/2)$.

so η_t is Z-distributed with $\eta_t \sim Z(\alpha = 1/2, \beta = 1/2, \mu_z = 0, \sigma_z = 1)$. Importantly, Z-distributions may be written as mean-variance scale mixtures of Gaussian distributions (Barndorff-Nielsen et al., 1982), which produces a useful framework for a parameter-expanded Gibbs sampler.

More generally, consider the inverted-Beta prior, denoted $IB(\beta, \alpha)$, for λ^2 with density

$$[\lambda^2] \propto \left(\lambda^2\right)^{\alpha-1}\left(1+\lambda^2\right)^{-(\alpha+\beta)}, \lambda > 0$$

(e.g., Armagan et al., 2011; Polson and Scott, 2012a,b). Special cases of the inverted-Beta distribution are provided in Table 3.1.

This broad class of priors may be equivalently constructed via the variances λ_t^2, the shrinkage parameters κ_t, or the log-variances η_t.

Proposition 3.1. *The following distributions are equivalent:*

1. $\lambda^2 \sim IB(\beta, \alpha)$;

2. $\kappa = 1/\left(1+\lambda^2\right) \sim$ *Beta*(β, α)*;*

3. $\eta = \log(\lambda^2) = \log(\kappa^{-1} - 1) \sim Z(\alpha, \beta, 0, 1)$.

Note that the ordering of the parameters α, β is identical for the inverted-Beta and Beta distributions, but reversed for the Z-distribution.

Now consider the dynamic setting in which $\phi \neq 0$. Model (3.2) implies that the conditional prior variance for ω_t in (3.1a) is $\exp(h_t) = \exp(\mu + \phi(h_{t-1} - \mu) + \eta_t) = \tau^2\lambda_{t-1}^{2\phi}\tilde{\lambda}_t^2$, where $\tau^2 = \exp(\mu)$, $\lambda_{t-1}^2 = \exp(h_{t-1} - \mu)$, and $\tilde{\lambda}_t^2 = \exp(\eta_t) \overset{iid}{\sim} IB(\beta, \alpha)$, as in the non-dynamic setting. This prior generalizes the $IB(\beta, \alpha)$ prior via the local variance term, $\lambda_{t-1}^{2\phi}$, which incorporates information about the shrinkage behavior at the previous time $t-1$ in the prior for ω_t. We formalize the role of this local adjustment term with the following results.

Proposition 3.2. *Suppose* $\eta \sim Z(\alpha, \beta, \mu_z, 1)$ *for* $\mu_z \in \mathbb{R}$. *Then* $\kappa = 1/(1+\exp(\eta)) \sim$ *TPB*$(\beta, \alpha, \exp(\mu_z))$*, where* $\kappa \sim$ *TPB*(β, α, γ) *denote the* three-parameter Beta distribution *with density* $[\kappa] = [B(\beta, \alpha)]^{-1}\gamma^{\beta}\kappa^{\beta-1}(1-\kappa)^{\alpha-1}\left[1 + (\gamma-1)\kappa\right]^{-(\alpha+\beta)}, \kappa \in (0,1), \gamma > 0$.

The three-parameter Beta (TPB) distribution generalizes the Beta distribution: $\gamma = 1$ produces the Beta(β, α) distribution, while $\gamma > 1$ (respectively, $\gamma < 1$) allocates more mass near zero (respectively, one) relative to the Beta(β, α) distribution. For dynamic shrinkage processes, the TPB distribution arises as the conditional prior distribution of κ_{t+1} given $\{\kappa_s\}_{s \leq t}$.

Theorem 3.1. *For the dynamic shrinkage process* (3.2)*, the conditional prior distribution of the shrinkage parameter* $\kappa_{t+1} = 1/\left(1+\tau^2\lambda_{t+1}^2\right)$ *is*

$$[\kappa_{t+1} | \{\kappa_s\}_{s \leq t}, \phi, \tau] \sim \textit{TPB}\left(\beta, \alpha, \tau^{2(1-\phi)}\left[\frac{1-\kappa_t}{\kappa_t}\right]^{\phi}\right) \tag{3.5}$$

or equivalently, $[\kappa_{t+1}|\{\lambda_s\}_{s\leq t}, \phi, \tau] \sim$ *TPB*$(\beta, \alpha, \tau^2\lambda_t^{2\phi})$.

The proof of Theorem 3.1 is in Appendix B. Naturally, the previous value of the shrinkage parameter, κ_t, together with the AR(1) coefficient ϕ, inform both the magnitude and the direction of the distributional shift of κ_{t+1}.

Theorem 3.2. *For the dynamic horseshoe process of* (3.2) *with* $\alpha = \beta = 1/2$ *and fixed* $\tau = 1$*, the conditional prior distribution* (3.5) *satisfies* $\mathbb{P}\big(\kappa_{t+1} < \varepsilon|\{\kappa_s\}_{s\leq t}, \phi\big) \to 1$ *as* $\kappa_t \to 0$ *for any* $\varepsilon \in (0, 1)$ *and fixed* $\phi \neq 0$*.*

The proof of Theorem 3.2 is in Appendix B. Importantly, Theorem 3.2 demonstrates that the mass of the conditional prior distribution for κ_{t+1} concentrates near zero—corresponding to minimal shrinkage of signals—when κ_t is near zero, so the shrinkage behavior at time t informs the (prior) shrinkage behavior at time $t + 1$.

We similarly characterize the posterior distribution of κ_{t+1} given $\{\kappa_s\}_{s\leq t}$ in the following theorem, which extends the results of Datta and Ghosh (2013) to the dynamic setting.

Theorem 3.3. *Under the likelihood* $y_t \overset{indep}{\sim} N(\omega_t, 1)$*, the prior* (3.1a)*, and the dynamic horseshoe process* (3.2) *with* $\alpha = \beta = 1/2$ *and fixed* $\phi \neq 0$*, the posterior distribution of* κ_{t+1} *given the history of the shrinkage process* $\{\kappa_s\}_{s\leq t}$ *satisfies the following properties:*

(a) For any fixed $\varepsilon \in (0, 1)$*,* $\mathbb{P}\big(\kappa_{t+1} > 1 - \varepsilon\big|y_{t+1}, \{\kappa_s\}_{s\leq t}, \phi, \tau\big) \to 1$ *as* $\gamma_t \to 0$ *uniformly in* $y_{t+1} \in \mathbb{R}$*, where* $\gamma_t = \tau^{2(1-\phi)}\left[(1 - \kappa_t)/\kappa_t\right]^{\phi}$*.*

(b) For any fixed $\varepsilon \in (0, 1)$ *and* $\gamma_t < 1$*,* $\mathbb{P}\big(\kappa_{t+1} < \varepsilon\big|y_{t+1}, \{\kappa_s\}_{s\leq t}, \phi, \tau\big) \to 1$ *as* $|y_{t+1}| \to \infty$*.*

The proof of Theorem 3.3 is in Appendix B, and uses the observation that marginally, $[y_{t+1}|\{\kappa_s\}] \overset{indep}{\sim} N(0, \kappa_{t+1}^{-1})$, so the posterior distribution of κ_{t+1} is

$$\begin{aligned}[\kappa_{t+1}|y_{t+1}, \{\kappa_s\}_{s\leq t}, \phi, \tau] &\propto \left\{\kappa_{t+1}^{\beta-1}(1-\kappa_{t+1})^{\alpha-1}\left[1+(\gamma_t-1)\kappa_{t+1}\right]^{-(\alpha+\beta)}\right\} \\ &\quad \times \left\{\kappa_{t+1}^{1/2}\exp\left(-y_{t+1}^2\kappa_{t+1}/2\right)\right\} \\ &\propto (1-\kappa_{t+1})^{-1/2}\left[1+(\gamma_t-1)\kappa_{t+1}\right]^{-1}\exp\left(-y_{t+1}^2\kappa_{t+1}/2\right).\end{aligned}$$

Theorem 3.3(a) demonstrates that the posterior mass of $[\kappa_{t+1}|\{\kappa_s\}_{s\leq t}]$ concentrates near one as $\tau \to 0$, as in the non-dynamic horseshoe, but also as $\kappa_t \to 1$. Therefore, the dynamic horseshoe process provides an additional mechanism for shrinkage of noise, besides the global scale parameter τ, via the previous shrinkage parameter κ_t. Moreover, Theorem 3.3(b) shows that, despite the additional shrinkage capabilities, the posterior mass of $[\kappa_{t+1}|\{\kappa_s\}_{s\leq t}]$ concentrates near zero for large absolute signals $|y_{t+1}|$, which indicates robustness of the dynamic horseshoe process to large signals analogous to the static horseshoe prior.

When $|\phi| < 1$, the log-volatility process $\{h_t\}$ is stationary, which implies $\{\kappa_t\}$ is stationary. In Figure 3.2, we plot a simulation-based estimate of the stationary distribution of κ_t for various values of ϕ under the dynamic horseshoe process. The stationary distribution of κ_t is similar to the static horseshoe distribution ($\phi = 0$) for $\phi < 0.5$, while for large values of ϕ the distribution becomes more peaked at zero (less shrinkage of ω_t) and one (more shrinkage of ω_t). The result is intuitive: larger $|\phi|$ corresponds to greater persistence in shrinkage behavior, so marginally we expect states of aggressive shrinkage or little shrinkage.

3.2.3 Scale Mixtures via Pólya-Gamma Processes

Standard SV sampling algorithms rely on a Gaussian assumption for the log-volatility innovations, $\eta_t \overset{iid}{\sim} N(0, \sigma_\eta^2)$, to efficiently sample the log-volatilities $\{h_t\}$ (e.g., Kim et al., 1998; Omori et al., 2007; Kastner and Frühwirth-Schnatter, 2014). To extend these techniques to the dynamic shrinkage process (3.2) in which $\eta_t \overset{iid}{\sim} Z(\alpha, \beta, 0, 1)$, we use parameter expansion to write η_t as a scale mixture of Gaussian distributions. The representation of a Z-distribution as a mean-variance scale mixtures of Gaussian distributions is due to Barndorff-Nielsen et al. (1982). For parameter expansion, we build on the framework of Polson et al. (2013), who propose a Pólya-Gamma scale mixture of Gaussians representation for Bayesian logistic regression. Importantly, this representation allows us to construct an efficient sampling algorithm that combines an $\mathcal{O}(T)$ sampling algorithm for the log-volatilities $\{h_t\}_{t=1}^T$ with a Pólya-Gamma sampler for the mixing parameters.

A *Pólya-Gamma* random variable ξ with parameters $b > 0$ and $c \in \mathbb{R}$, denoted $\xi \sim \text{PG}(b, c)$, is an infinite convolution of Gamma random variables:

$$\xi \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 - c^2/(4\pi^2)} \tag{3.6}$$

where $g_k \overset{iid}{\sim} \text{Gamma}(b, 1)$. Properties of Pólya-Gamma random variables may be found in Barndorff-Nielsen et al. (1982) and Polson et al. (2013). Our interest in Pólya-Gamma random variables derives from their role in representing the Z-distribution as a mean-variance scale mixture of Gaussians.

Theorem 3.4. *The random variable* $\eta \sim Z(\alpha, \beta, 0, 1)$*, or equivalently* $\eta = \log(\lambda^2)$

with $\lambda^2 \sim IB(\beta, \alpha)$, is a mean-variance scale mixture of Gaussian distributions with

$$\begin{cases} [\eta|\xi] \sim N\left(\xi^{-1}[\alpha - \beta]/2, \xi^{-1}\right) \\ [\xi] \sim PG(\alpha + \beta, 0). \end{cases} \tag{3.7}$$

Moreover, the conditional distribution of ξ is $[\xi|\eta] \sim PG(\alpha + \beta, \eta)$.

The proof of Theorem 3.4 is in Appendix B. When $\alpha = \beta$, the Z-distribution is symmetric, and the conditional expectation in (3.7) simplifies to $\mathbb{E}[\eta|\xi] = 0$. Polson et al. (2013) propose a sampling algorithm for Pólya-Gamma random variables, which is available in the `R` package `BayesLogit`, and is extremely efficient when $b = 1$. In our setting, this corresponds to $\alpha + \beta = 1$, for which the horseshoe prior is the prime example.

3.3 Bayesian Trend Filtering with Dynamic Shrinkage Processes

Dynamic shrinkage processes are particularly appropriate for dynamic linear models (DLMs). DLMs combine an observation equation, which relates the observed data to latent state variables, and an evolution equation, which allows the state variables—and therefore the conditional mean of the data—to be dynamic. By construction, DLMs contain many parameters, and therefore may benefit from structured regularization. The proposed dynamic shrinkage processes offer such regularization, and unlike existing methods, do so adaptively.

Consider the following DLM with a Dth order random walk on the state

variable, β_t:

$$\begin{cases} y_t = \beta_t + \epsilon_t, & [\epsilon_t|\sigma_\epsilon] \overset{iid}{\sim} N(0, \sigma_\epsilon^2), \quad t = 1, \ldots, T \\ \Delta^D \beta_{t+1} = \omega_t, & [\omega_t|\tau, \{\lambda_s\}] \overset{indep}{\sim} N(0, \tau^2\lambda_t^2), \quad t = D, \ldots, T \end{cases} \tag{3.8}$$

and $\beta_{t+1} = \omega_t \sim N(0, \tau^2\lambda_t^2)$ for $t = 0, \ldots, D-1$, where Δ is the differencing operator and $D \in \mathbb{Z}^+$ is the degree of differencing. By imposing a shrinkage prior on λ_t, model (3.8) may be viewed as a Bayesian adaptation of the *trend filtering* model of Kim et al. (2009) and Tibshirani (2014): model (3.8) features a penalty encouraging sparsity of the Dth order differences of the conditional mean, β_t. Faulkner and Minin (2016) provide an implementation based on the (static) horseshoe prior and the Bayesian lasso, and further allow for non-Gaussian likelihoods. We refer to model (3.8) as a *Bayesian trend filtering* (BTF) model, with various choices available for the distribution of the innovation standard deviations, $[\tau\lambda_t]$.

We propose a dynamic horseshoe process as the prior for the innovations ω_t in model (3.8). The aggressive shrinkage of the horseshoe prior forces small values of $|\omega_t| = |\Delta^D \beta_{t+1}|$ toward zero, while the robustness of the horseshoe prior permits large values of $|\Delta^D \beta_{t+1}|$. When $D = 2$, model (3.8) will shrink the conditional mean β_t toward a piecewise linear function with breakpoints determined adaptively, while allowing large absolute changes in the slopes. Further, using the *dynamic* horseshoe process, the shrinkage effects induced by λ_t are time-dependent, which provides localized adaptability to regions with rapidly- or slowly-changing features. Following Carvalho et al. (2010) and Polson and Scott (2012b), we assume a half-Cauchy prior for the global scale parameter $\tau \sim C^+(0, \sigma_\epsilon/\sqrt{T})$, in which we scale by the observation error variance and the sample size (Piironen and Vehtari, 2016). Using Pólya-Gamma mixtures, the implied conditional prior on $\mu = \log(\tau^2)$ is $[\mu|\sigma_\epsilon, \xi_\mu] \sim N(\log \sigma_\epsilon^2 - \log T, \xi_\mu^{-1})$

with $\xi_\mu \sim \text{PG}(1, 0)$. We include the details of the Gibbs sampling algorithm for model (3.8) in Section 3.5, which is notably *linear* in the number of time points, T: the full conditional posterior precision matrices for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_T)'$ and $\boldsymbol{h} = (h_1, \ldots, h_T)'$ are D-banded and tridiagonal, respectively, which admit highly efficient $\mathcal{O}(T)$ back-band substitution sampling algorithms (see Appendix B for empirical evidence).

3.3.1 Bayesian Trend Filtering: Simulations

To assess the performance of the Bayesian trend filtering (BTF) model (3.8) with dynamic horseshoe innovations (**BTF-DHS**), we compared the proposed methods to several competitive alternatives using simulated data. We considered the following variations on BTF model (3.8): normal-inverse-Gamma (**BTF-NIG**) innovations via $\tau^{-2} \sim \text{Gamma}(0.001, 0.001)$ with $\lambda_t = 1$; and (static) horseshoe priors for the innovations (**BTF-HS**) via $\tau, \lambda_t \overset{iid}{\sim} C^+(0, 1)$. In addition, we include the (non-Bayesian) trend filtering model of Tibshirani (2014) implemented using the `R` package `genlasso` (Arnold and Tibshirani, 2014), for which the regularization tuning parameter is chosen using cross-validation (**Trend Filtering**). For all trend filtering models, we select $D = 2$, but the relative performance is similar for $D = 1$. Among non-trend filtering models, we include a smoothing spline estimator implemented via `smooth.spline()` in `R` (**Smoothing Spline**); the wavelet-based estimator of Abramovich et al. (1998) (**BayesThresh**) implemented in the `wavethresh` package (Nason, 2016); and the nested Gaussian Process (**nGP**) model of Zhu and Dunson (2013), which relies on a state space model framework for efficient computations, comparable to—but empirically less efficient than—the BTF model (3.8).

We simulated 100 data sets from the model $y_t = y_t^* + \epsilon_t$, where y_t^* is the true function and $\epsilon_t \overset{indep}{\sim} N(0, \sigma_*^2)$. We use the following true functions y_t^* from Donoho and Johnstone (1994): *Doppler*, *Bumps*, *Blocks*, and *Heavisine*, implemented in the `R` package `wmtsa` (Constantine and Percival, 2016). The noise variance σ_*^2 is determined by selecting a root-signal-to-noise ratio (RSNR) and computing $\sigma_* = \sqrt{\frac{\sum_{t=1}^{T}(y_t^* - \bar{y}^*)^2}{T-1}} \Big/ \text{RSNR}$, where $\bar{y}^* = \frac{1}{T}\sum_{t=1}^{T} y_t^*$. As in Zhu and Dunson (2013), we select RSNR = 7 and use a moderate length time series, $T = 128$.

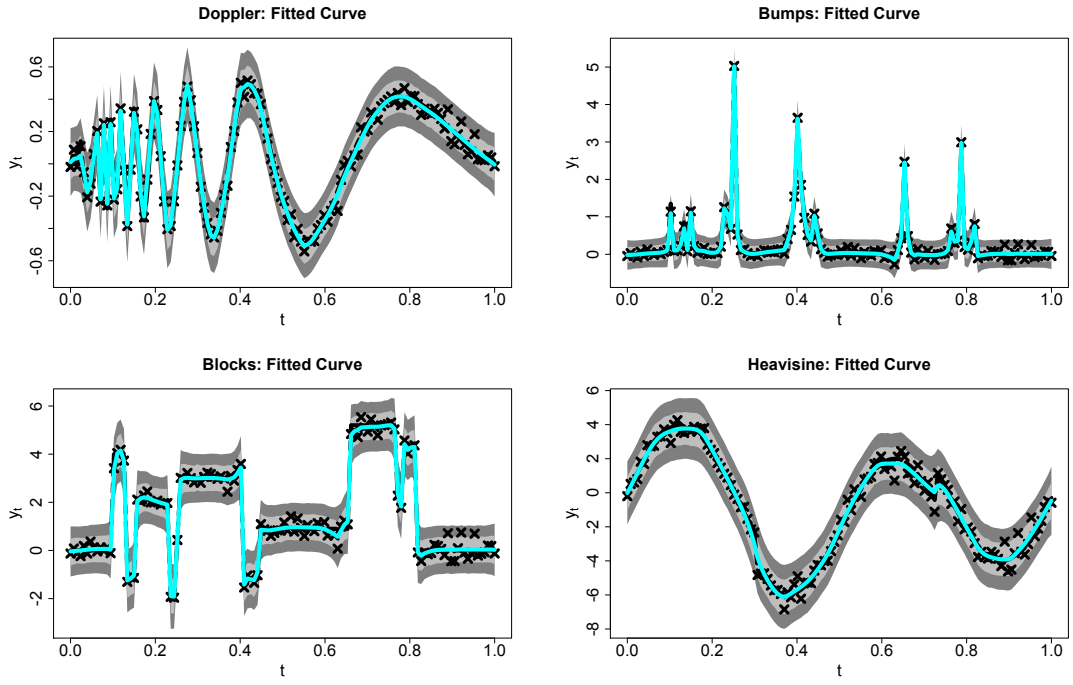


Figure 3.3: Fitted curves for simulated data with $T = 128$ and RSNR $= 7$. Each panel includes the simulated observations (x-marks), the posterior expectations of β_t (cyan), and the 95% pointwise HPD credible intervals (light gray) and 95% simultaneous credible bands (dark gray) for the posterior predictive distribution of $\{y_t\}$ under BTF-DHS model (3.8) with $D = 2$. The proposed estimator, as well as the uncertainty bands, accurately capture both slowly- and rapidly-changing behavior in the underlying functions.

In Figure 3.3, we provide an example of each true curve y_t^*, together with the proposed BTF-DHS posterior expectations and credible bands. Notably, the

Bayesian trend filtering model (3.8) with $D = 2$ and dynamic horseshoe innovations provides an exceptionally accurate fit to each data set. Importantly, the posterior expectations and the posterior credible bands adapt to both slowly- and rapidly-changing behavior in the underlying curves. The implementation is also efficient: the computation time for 15,000 iterations of the Gibbs sampling algorithm, implemented in R (on a MacBook Pro, 2.7 GHz Intel Core i5), is about 1.15 minutes.

To compare the aforementioned procedures, we compute the root mean squared errors $\text{RMSE}(\hat{y}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^* - \hat{y}_t)^2}$ for all estimators $\hat{y}$ of the true function, y^*. The results are displayed in Figure 3.4. The proposed BTF-DHS implementation substantially outperforms all competitors, especially for rapidly-changing curves (Doppler and Bumps). The exceptional performance of BTF-DHS is paired with comparably small variability of RMSE, especially relative to non-dynamic horseshoe model (BTF-HS). Interestingly, the magnitude and variability of the RMSEs for BTF-DHS are related to the AR(1) coefficient, ϕ: the 95% HPD intervals (corresponding to Figure 3.3) are $(0.77, 0.97)$ (Doppler), $(0.81, 0.97)$ (Bumps), $(0.76, 0.96)$ (Blocks), and $(-0.04, 0.74)$ (Heavisine). For the smoothest function, Heavisine, there is less separation among the estimators. Nonetheless, BTF-DHS performs the best, even though the HPD interval for ϕ is wider and contains zero. These results show that the Bayesian trend filtering model (3.8) with dynamic horseshoe innovations substantially improves upon existing curve-fitting procedures, and due to both its computational efficiency and the availability of posterior inference, may provide a useful procedure for a wide variety of applications.

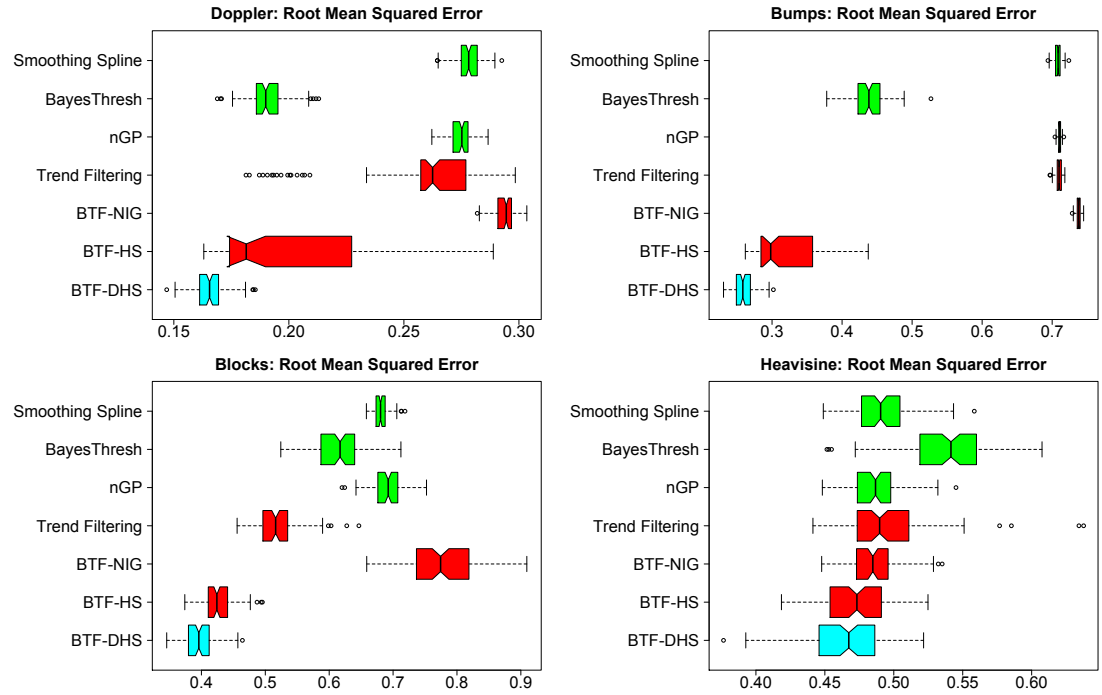


Figure 3.4: Root mean squared errors for simulated data with $T = 128$ and RSNR $= 7$. The Bayesian trend filtering (BTF) estimators differ in their innovation distributions, which determines the shrinkage behavior of the second order differences ($D = 2$): normal-inverse-Gamma (NIG), horseshoe (HS), and dynamic horseshoe (DHS).

3.3.2 Bayesian Trend Filtering: Application to CPU Usage Data

To demonstrate the adaptability of the dynamic horseshoe process for model (3.8), we consider the CPU usage data in Figure 3.1a. The data exhibit substantial complexity: an overall smooth intraday trend but with multiple irregularly-spaced jumps, and an increase in volatility from 16:00-18:00. Our goal is to provide an accurate measure of the trend, including jumps, with appropriate uncertainty quantification. For this purpose, we employ the BTF-DHS model (3.8), which we extend to include stochastic volatility for the observation error: $y_t \overset{indep}{\sim} N(\beta_t, \sigma_t^2)$ with an AR(1) model on $\log \sigma_t^2$ as in (3.2) with $\eta_t \overset{iid}{\sim} N(0, \sigma_\eta^2)$. For the additional sampling step of the stochastic volatility parameters, we use

the algorithm of Kastner and Frühwirth-Schnatter (2014) implemented in the `R` package `stochvol` (Kastner, 2016).

The resulting model fit is summarized in Figure 3.1. The posterior expectation and posterior credible bands accurately model both irregular jumps and smooth trends, and capture the increase in volatility from 16:00-18:00 (see Figure 3.1c). By examining regions of nonoverlapping simultaneous posterior credible bands, we may assess change points in the level of the data. In particular, the model fit suggests that the CPU usage followed a slowly increasing trend interrupted by jumps of two distinct magnitudes prior to 16:00, after which the volatility increased and the level decreased until approximately 18:00.

We augment the simulation study of Section 3.3.1 with a comparison of out-of-sample estimation of the CPU usage data. We fit each model using 90% ($T = 1296$) of the data selected randomly for training and the remaining 10% ($T = 144$) for testing, which was repeated independently 100 times. Models were compared using RMSE.

Unlike the simulation study in Section 3.3.1, the subsampled data are *not* equally spaced. Taking advantage of the computational efficiency of the proposed BTF methodology, we employ a model-based imputation scheme, which is valid for missing observations. For unequally-spaced data $y_{t_i}, i = 1, \ldots, T$, we expand the operative data set to include missing observations along an equally-spaced grid, $t^* = 1, \ldots, T^*$, such that for each observation point i, $y_{t_i} = y_{t^*}$ for some t^*. Although $T^* \geq T$, possibly with $T^* \gg T$, all computations within the sampling algorithm, including the imputation sampling scheme for $\{y_{t^*} : t^* \neq t_i\}$, are linear in the number of (equally-spaced) time points, T^*. Therefore, we may apply the same Gibbs sampling algorithm as before, with

the additional step of drawing $y_{t^*} \overset{indep}{\sim} N(\beta_{t^*}, \sigma^2_{t^*})$ for each unobserved $t^* \neq t_i$. Implicitly, this procedure assumes that the unobserved points are missing at random, which is satisfied by the aforementioned subsampling scheme.

The results of the out-of-sample estimation study are displayed in Figure 3.5. The BTF procedures are notably superior to the non-Bayesian trend filtering and smoothing spline estimators, and, as with the simulations of Section 3.3.1, the proposed BTF-DHS model substantially outperforms all competitors.

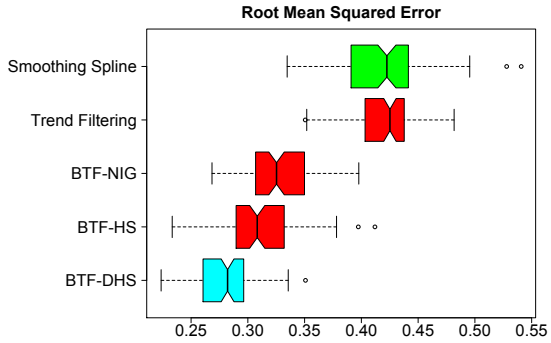


Figure 3.5: Root mean squared error for out-of-sample minute-by-minute CPU usage data. The Bayesian trend filtering (BTF) estimators differ in their innovation distributions, which determines the shrinkage behavior of the second order differences ($D = 2$): normal-inverse-Gamma (NIG), horseshoe (HS), and dynamic horseshoe (DHS).

3.4 Joint Shrinkage for Time-Varying Parameter Models

Dynamic shrinkage processes are appropriate for multivariate time series models that may benefit from locally adaptive shrinkage properties. Consider the following time-varying parameter regression model with multiple dynamic predictors $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{p,t})'$:

$$\begin{cases} y_t = \boldsymbol{x}_t' \boldsymbol{\beta}_t + \epsilon_t, & [\epsilon_t | \sigma_\epsilon] \overset{indep}{\sim} N(0, \sigma^2_\epsilon) \\ \Delta^D \boldsymbol{\beta}_{t+1} = \boldsymbol{\omega}_t, & [\omega_{j,t} | \tau_0, \{\tau_k\}, \{\lambda_{k,s}\}] \overset{indep}{\sim} N(0, \tau_0^2 \tau_j^2 \lambda_{j,t}^2) \end{cases} \tag{3.9}$$

where $\boldsymbol{\beta}_t = (\beta_{1,t}, \ldots, \beta_{p,t})'$ is the vector of dynamic regression coefficients and $D \in \mathbb{Z}^+$ is the degree of differencing. The prior for the innovations $\omega_{j,t}$ incorporates three levels of global-local shrinkage: a global shrinkage parameter τ_0, a predictor-specific shrinkage parameter τ_j, and a predictor- and time-specific local shrinkage parameter $\lambda_{j,t}$.

To provide jointly localized shrinkage of the dynamic regression coefficients $\{\beta_{j,t}\}$ analogous to the Bayesian trend filtering model of Section 3.3, we extend (3.2) to allow for multivariate time dependence via a vector autoregression (VAR) on the log-volatility:

$$\begin{cases} [\omega_{j,t} | \tau_0, \{\tau_k\}, \{\lambda_{k,s}\}] \overset{indep}{\sim} N(0, \tau_0^2 \tau_j^2 \lambda_{j,t}^2) & \\ h_{j,t} = \log(\tau_0^2 \tau_j^2 \lambda_{j,t}^2), & j = 1, \ldots, p, t = 1, \ldots, T \\ \boldsymbol{h}_{t+1} = \boldsymbol{\mu} + \boldsymbol{\Phi}(\boldsymbol{h}_t - \boldsymbol{\mu}) + \boldsymbol{\eta}_t, & \eta_{j,t} \overset{iid}{\sim} Z(\alpha, \beta, 0, 1) \end{cases} \tag{3.10}$$

where $\boldsymbol{h}_t = (h_{1,t}, \ldots, h_{p,t})'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, $\boldsymbol{\eta}_t = (\eta_{1,t}, \ldots, \eta_{p,t})'$, and $\boldsymbol{\Phi}$ is the $p \times p$ VAR coefficient matrix. We assume $\boldsymbol{\Phi} = \text{diag}\,(\phi_1, \ldots, \phi_p)$ for simplicity, but non-diagonal extensions are available. As in the univariate setting, we use Pólya-Gamma mixtures (independently) for the log-volatility evolution errors, $[\eta_{j,t} | \xi_{j,t}] \overset{indep}{\sim} N(\xi_{j,t}^{-1}[\alpha - \beta]/2, \xi_{j,t}^{-1})$ with $\xi_{j,t} \overset{iid}{\sim} \text{PG}(\alpha + \beta, 0)$ and $\alpha = \beta = 1/2$. We augment model (3.10) with half-Cauchy priors for the predictor-specific and global parameters, $\tau_j \overset{indep}{\sim} C^+(0, 1)$ and $\tau_0 \sim C^+(0, \sigma_\epsilon/\sqrt{Tp})$, in which we scale by the observation error variance and the number of innovations $\{\omega_{j,t}\}$ (Piironen and Vehtari, 2016). These priors may be equivalently represented on the log-scale using the Pólya-Gamma parameter expansion $[\mu_j | \mu, \xi_{\mu_j}] \sim N(\mu, \xi_{\mu_j}^{-1})$ and $[\mu_0 | \sigma_\epsilon, \xi_{\mu_0}] \sim N(\log \sigma_\epsilon^2 - \log T, \xi_{\mu_0}^{-1})$ with $\xi_{\mu_j}, \xi_{\mu_0} \overset{iid}{\sim} \text{PG}(1, 0)$ and the identification $\mu_j = \log(\tau_0^2 \tau_j^2)$ and $\mu_0 = \log(\tau_0^2)$.

3.4.1 Time-Varying Parameter Models: Simulations

We conducted a simulation study to evaluate competing variations of the time-varying parameter regression model (3.9), in particular relative to the proposed dynamic shrinkage process (**BTF-DHS**) in (3.10). Similar to the simulations of Section 3.3.1, we focus on the distribution of the innovations, $\omega_{j,t}$, and again include the normal-inverse-Gamma (**BTF-NIG**) and the (static) horseshoe (**BTF-HS**) as competitors, in each case selecting $D = 2$. Among models with non-dynamic regression coefficients, we include a lasso regression (Tibshirani, 1996) implemented via the `R` package `glmnet` (Friedman et al., 2010), which incorporates variable selection, and an ordinary linear regression.

We simulated 100 data sets of length $T = 500$ from the model $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_t^* + \epsilon_t$, where the $p = 7$ predictors are $x_{1,t} = 1$ and $x_{j,t} \overset{iid}{\sim} N(0,1)$ for $j > 2$, and $\epsilon_t \overset{iid}{\sim} N(0, \sigma_*^2)$. The true regression coefficients $\boldsymbol{\beta}_t^* = (\beta_{1,t}^*, \ldots, \beta_{p,t}^*)'$ are the following: $\beta_{1,t}^* = 1$, $\beta_{2,t}^*$ and $\beta_{3,t}^*$ are the *Bumps* and *Heavisine* functions, respectively, from Section 3.3.1 rescaled to $[0,1]$, and $\beta_{j,t}^* = 0$ for $j = 4, \ldots, p = 7$. The predictor set contains a variety of functions: a constant nonzero function, a rapidly-changing function (Bumps), a relatively smooth function (Heavisine), and three true zeros. The noise variance σ_*^2 is determined by selecting a root-signal-to-noise ratio (RSNR) and computing $\sigma_* = \sqrt{\frac{\sum_{t=1}^T (y_t^* - \bar{y}^*)^2}{T-1}} \Big/ \text{RSNR}$, where $y_t^* = \boldsymbol{x}_t'\boldsymbol{\beta}_t^*$ and $\bar{y}^* = \frac{1}{T}\sum_{t=1}^T y_t^*$. We select RSNR $= 10$.

In Figure 3.6, we show the true regression functions $\beta_{j,t}^*$, together with the proposed BTF-DHS posterior expectations and credible bands for $\beta_{j,t}$. Despite the challenge presented by the Bumps function, the proposed model (3.9) with innovation distribution (3.10) adequately identifies the constant and zero curves, captures the important features of the Bumps function, and accurately

estimates the smoother Heavisine function.

We evaluate competing methods using RMSEs for both y_t^* and $\boldsymbol{\beta}_t^*$ defined by $\text{RMSE}(\hat{y}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t^* - \hat{y}_t)^2}$ and $\text{RMSE}(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{Tp}\sum_{t=1}^{T}\sum_{j=1}^{p}\left(\beta_{j,t}^* - \hat{\beta}_{j,t}\right)^2}$ for all estimators $\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}$ of the true regression functions, $\boldsymbol{\beta}_t^*$ with $\hat{y}_t = \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}$. The results are displayed in Figure 3.7. The proposed BTF-DHS model substantially outperforms the competitors in both recovery of the true regression functions, $\beta_{j,t}^*$ and estimation of the true curves, y_t^*. Notably, the dynamic (BTF) procedures offer massive gains over the models with static regression coefficients.

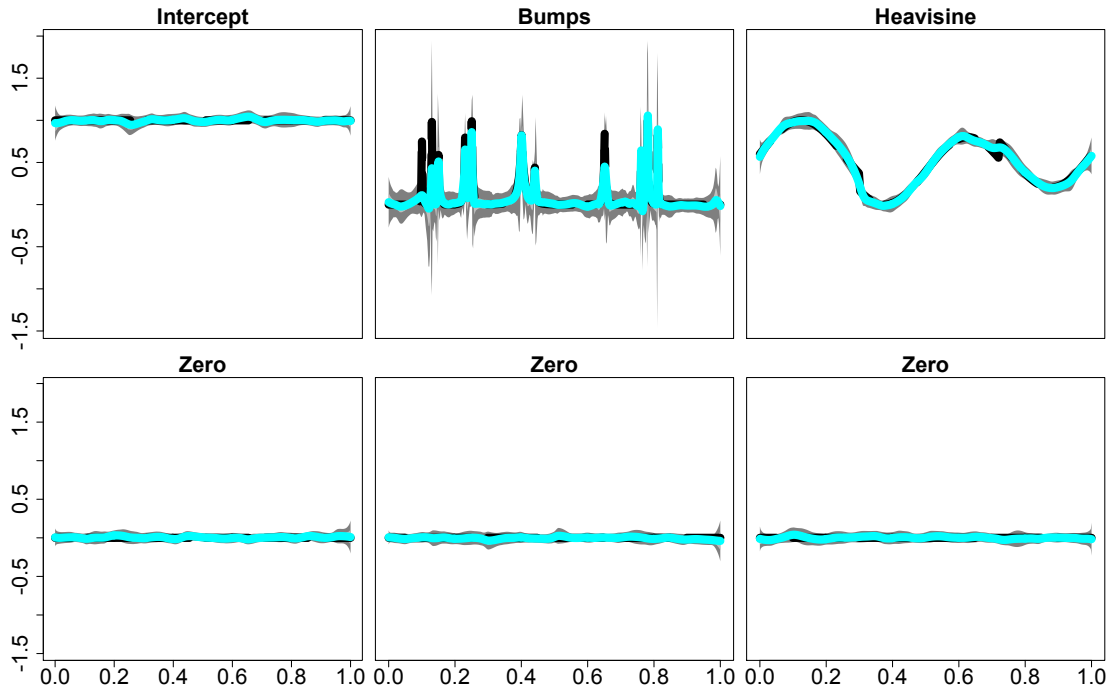


Figure 3.6: True regression functions $\beta_{j,t}^*$ (black line) and corresponding posterior expectations (cyan), 95% pointwise HPD credible intervals (light gray) and 95% simultaneous credible bands (dark gray) for $\beta_{j,t}$ under the BTF-DHS model given by (3.9) and (3.10) for a simulated data set.

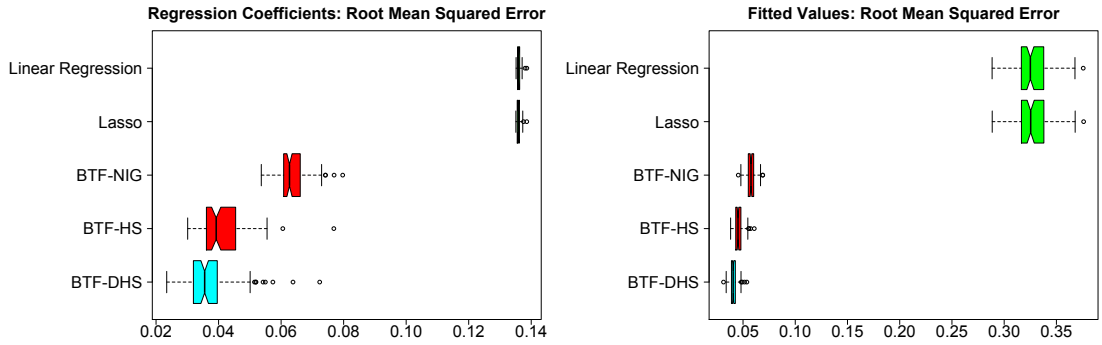


Figure 3.7: Root mean squared errors for the regression coefficients, $\beta^*_{j,t}$ (**left**) and the true curves, $y^*_t = \boldsymbol{x}'_t\boldsymbol{\beta}^*_t$ (**right**) for simulated data.

3.4.2 Time-Varying Parameter Models: The Fama-French Asset Pricing Model

Asset pricing models commonly feature highly structured factor models to parsimoniously model the co-movement of stock returns. Such fundamental factor models identify common risk factors among assets, which may be treated as exogenous predictors in a time series regression. Popular approaches include the one-factor Capital Asset Pricing Model (CAPM, Sharpe, 1964) and the three-factor Fama-French model (FF-3, Fama and French, 1993). Recently, the five-factor Fama-French model (FF-5, Fama and French, 2015) was proposed as an extension of FF-3 to incorporate additional common risk factors. However, outstanding questions remain regarding which, and how many, factors are necessary. Importantly, an attempt to address these questions must consider the dynamic component: the relevance of individual factors may change over time, particularly for different assets.

We apply model (3.9) to extend these fundamental factor models to the dynamic setting, in which the factor loadings are permitted to vary—perhaps

rapidly—over time. For further generality, we append the momentum factor of Carhart (1997) to FF-5 to produce a fundamental factor model with six factors and dynamic factor loadings. Importantly, the shrinkage towards sparsity induced by the dynamic horseshoe process allows the model to effectively select out unimportant factors, which also may change over time. As in Section 3.3.2, we modify model (3.9) to include stochastic volatility for the observation error, $[\epsilon_t|\{\sigma_s\}] \overset{indep}{\sim} N(0, \sigma_t^2)$.

To study various market sectors, we use weekly industry portfolio data from the website of Kenneth R. French, which provide the value-weighted return of stocks in the given industry. We focus on manufacturing (Manuf) and healthcare (Hlth). For a given industry portfolio, the response variable is the returns in excess of the risk free rate, $y_t = R_t - R_{F,t}$, with predictors $\boldsymbol{x}_t = (1, R_{M,t} - R_{F,t}, SMB_t, HML_t, RMW_t, CMA_t, MOM_t)'$, defined as follows: the *market risk factor*, $R_{M,t} - R_{F,t}$ is the return on the market portfolio $R_{M,t}$ in excess of the risk free rate $R_{F,t}$; the *size factor*, SMB_t (small minus big) is the difference in returns between portfolios of small and large market value stocks; the *value factor*, HML_t (high minus low) is the difference in returns between portfolios of high and low book-to-market value stocks; the *profitability factor*, RMW_t is the difference in returns between portfolios of robust and weak profitability stocks; the *investment factor*, CMA_t is the difference in returns between portfolios of stocks of low and high investment firms; and the *momentum factor*, MOM_t is the difference in returns between portfolios of stocks with high and low prior returns. These data are publicly available on Kenneth R. French's website, which provides additional details on the portfolios. We standardize all predictors and the response to have unit variance.

In Figures 3.8 and 3.9, we plot the posterior expectation and credible bands for the time-varying regression coefficients and observation error stochastic volatility for the weekly manufacturing and healthcare industry data sets, respectively, from 4/1/2007 - 4/1/2017 ($T = 522$). The 95% simultaneous credible bands (dark gray) indicate which coefficients are significantly different from zero, and if so, at which times. For the manufacturing industry, the significant factors are the market risk ($R_{M,t} - R_{F,t}$), investment (CMA_t), and momentum (MOM_t), where both CMA_t and MOM_t are significantly time-varying (i.e., the simultaneous credible bands contain no constant function). By comparison, an ordinary linear regression does *not* find MOM_t to be significant at the 5% level, since the non-dynamic model ignores the fluctuations from 2008-2012, but does identify the market risk, profitability (RMW_t), and investment as significant factors (see Appendix B for details).

For the healthcare industry, the significant factors are market risk, value (HML_t), and profitability. By comparison, the ordinary linear regression identifies these factors as well as size (SMB_t) as significant at the 5% level (see the Appendix B for details). Notably, the only common factor significant in both the manufacturing and healthcare industries under model (3.9) over this time period is the market risk. This result suggests that the aggressive shrinkage behavior of the dynamic shrinkage process is important in this setting, since several factors may be effectively irrelevant for some or all time points.

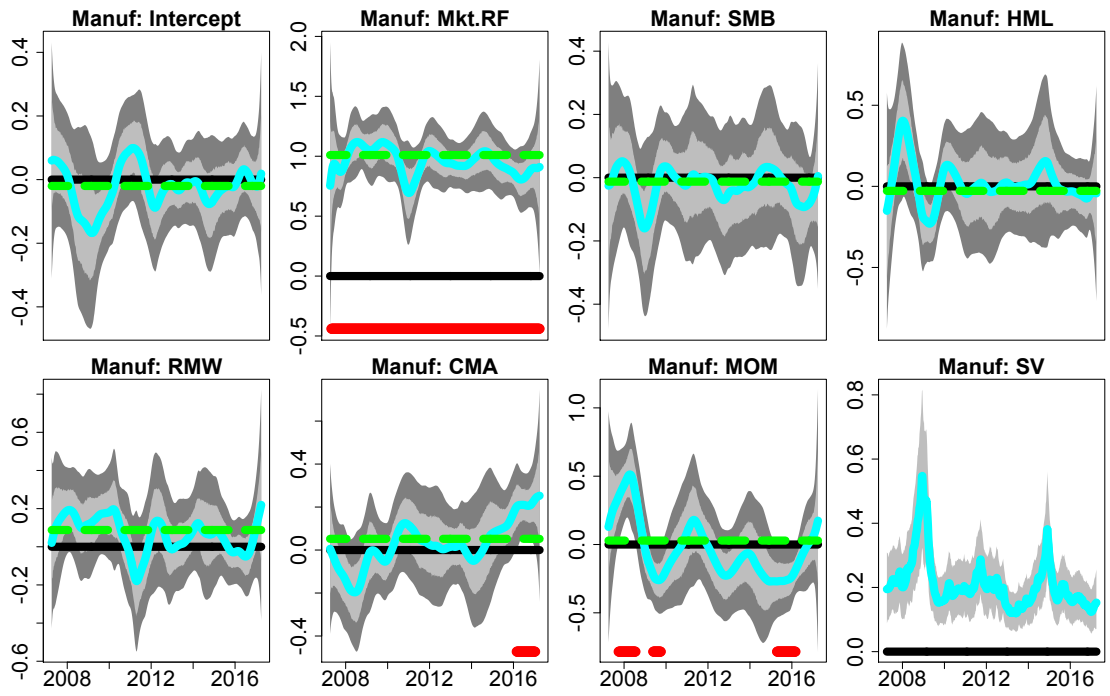


Figure 3.8: Posterior expectations (cyan), 95% pointwise HPD credible intervals (light gray) and 95% simultaneous credible bands (dark gray) for $\beta_{j,t}$ and σ_t (bottom right) under the BTF-DHS model given by (3.9) and (3.10) for value-weighted manufacturing industry returns. The solid black line is zero, the dashed green line is the ordinary linear regression estimate, and the solid red line indicates periods for which the 95% simultaneous credible bands do not contain zero.

3.5 MCMC Sampling Algorithm and Computational Details

We design a Gibbs sampling algorithm for the dynamic shrinkage process. The sampling algorithm is both computationally and MCMC efficient, and builds upon two main components: (1) a stochastic volatility sampling algorithm (Kastner and Frühwirth-Schnatter, 2014) augmented with a Pólya-Gamma sampler (Polson et al., 2013); and (2) a Cholesky Factor Algorithm (CFA, Rue, 2001) for sampling the state variables in the dynamic linear model. Importantly, both components employ algorithms that are linear in the number of time points, which produces a highly efficient sampling algorithm.

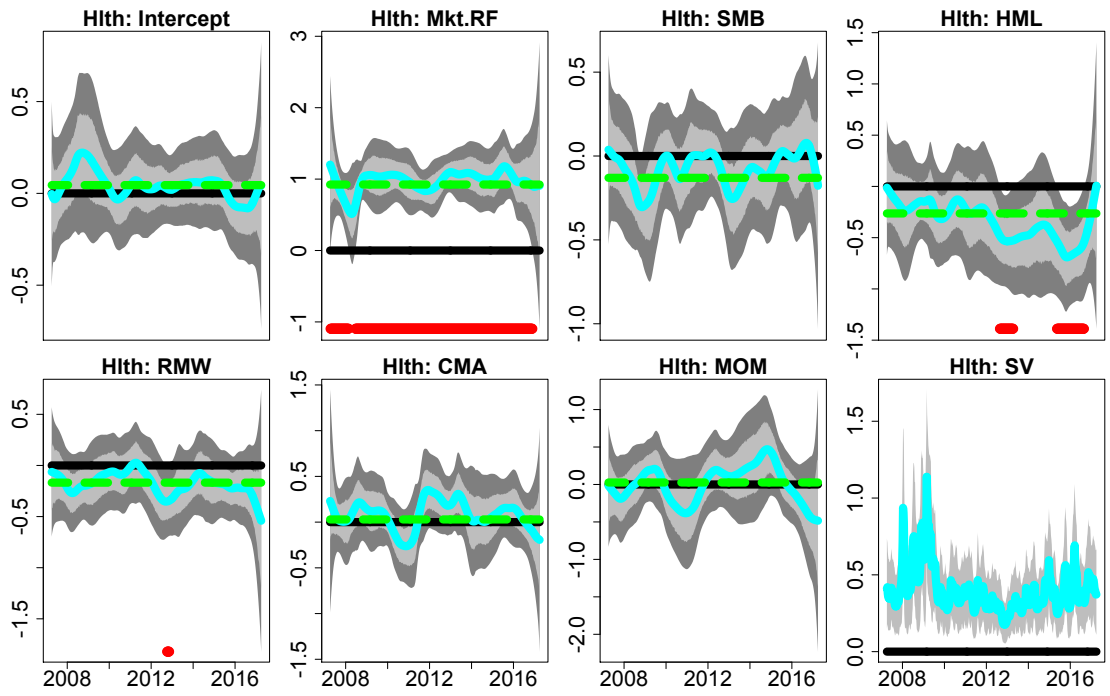


Figure 3.9: Posterior expectations (cyan), 95% pointwise HPD credible intervals (light gray) and 95% simultaneous credible bands (dark gray) for $\beta_{j,t}$ and σ_t (bottom right) under the BTF-DHS model given by (3.9) and (3.10) for value-weighted healthcare industry returns. The solid black line is zero, the dashed green line is the ordinary linear regression estimate, and the solid red line indicates periods for which the 95% simultaneous credible bands do not contain zero.

The general sampling algorithm is as follows: (1) sample the dynamic shrinkage components (the log-volatilities $\{h_t\}$, the Pólya-Gamma mixing parameters $\{\xi_t\}$, the unconditional mean of log-volatility μ, the AR(1) coefficient of log-volatility ϕ, and the discrete mixture component indicators $\{s_t\}$); (2) sample the state variables $\{\boldsymbol{\beta}_t\}$; and (3) sample the observation error variance σ_ϵ^2. We provide details of the dynamic shrinkage process sampling algorithm in Section 3.5.1 and include the details for sampling steps (2) and (3) in Appendix B.

3.5.1 Efficient Sampling for the Dynamic Shrinkage Process

Consider the (univariate) dynamic shrinkage process in (3.2) with the Pólya-Gamma parameter expansion of Theorem 3.4. We provide implementation details for the dynamic horseshoe process with $\alpha = \beta = 1/2$, but extensions to other cases are straightforward. The SV sampling framework of Kastner and Frühwirth-Schnatter (2014) represents the likelihood for h_t on the log-scale, and approximates the ensuing $\log \chi_1^2$ distribution for the errors via a known discrete mixture of Gaussian distributions. In particular, let $\tilde{y}_t = \log(\omega_t^2 + c)$, where c is a small offset to avoid numerical issues. Conditional on the mixture component indicators s_t, the likelihood is $\tilde{y}_t \overset{indep}{\sim} N(h_t + m_{s_t}, v_{s_t})$ where m_i and $v_i, i = 1, \ldots, 10$ are the pre-specified mean and variance components of the 10-component Gaussian mixture provided in Omori et al. (2007). The evolution equation is $h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t$ with initialization $h_1 = \mu + \eta_0$ and innovations $[\eta_t | \xi_t] \overset{indep}{\sim} N(0, \xi_t^{-1})$ for $[\xi_t] \overset{iid}{\sim} \text{PG}(1, 0)$.

To sample $\boldsymbol{h} = (h_1, \ldots, h_T)$ jointly, we directly compute the posterior distribution of $\boldsymbol{h}$ and exploit the tridiagonal structure of the resulting posterior precision matrix. In particular, we equivalently have $\tilde{\boldsymbol{y}} \sim N(\boldsymbol{m} + \tilde{\boldsymbol{h}} + \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_v)$ and $\boldsymbol{D}_\phi \tilde{\boldsymbol{h}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\xi)$, where $\boldsymbol{m} = (m_{s_1}, \ldots, m_{s_T})'$, $\tilde{\boldsymbol{h}} = (h_1 - \mu, \ldots, h_T - \mu)'$, $\tilde{\boldsymbol{\mu}} = (\mu, (1-\phi)\mu, \ldots, (1-\phi)\mu)'$, $\boldsymbol{\Sigma}_v = \text{diag}\left(\{v_{s_t}\}_{t=1}^T\right)$, $\boldsymbol{\Sigma}_\xi = \text{diag}\left(\{\xi_t^{-1}\}_{t=1}^T\right)$, and $\boldsymbol{D}_\phi$ is a lower triangular matrix with ones on the diagonal, $-\phi$ on the first off-diagonal, and zeros elsewhere. We sample from the posterior distribution of $\boldsymbol{h}$ by sampling from the posterior distribution of $\tilde{\boldsymbol{h}}$ and setting $\boldsymbol{h} = \tilde{\boldsymbol{h}} + \mu \boldsymbol{1}$ for $\boldsymbol{1}$ a T-dimensional vector of ones. The required posterior distribution is $\tilde{\boldsymbol{h}} \sim N\left(\boldsymbol{Q}_{\tilde{h}}^{-1} \boldsymbol{\ell}_{\tilde{h}}, \boldsymbol{Q}_{\tilde{h}}^{-1}\right)$, where $\boldsymbol{Q}_{\tilde{h}} = \boldsymbol{\Sigma}_v^{-1} + \boldsymbol{D}_\phi' \boldsymbol{\Sigma}_\xi^{-1} \boldsymbol{D}_\phi$ is a tridiagonal symmetric matrix with diagonal elements $\boldsymbol{d}_0(\boldsymbol{Q}_{\tilde{h}})$ and first off-diagonal elements $\boldsymbol{d}_1(\boldsymbol{Q}_{\tilde{h}})$

defined as

$$\boldsymbol{d}_0(\boldsymbol{Q}_{\tilde{h}}) = \Big[(v_{s_1}^{-1} + \xi_1 + \phi^2\xi_2), (v_{s_2}^{-1} + \xi_2 + \phi^2\xi_3), \ldots, (v_{s_{T-1}}^{-1} + \xi_{T-1} + \phi^2\xi_T), (v_{s_T}^{-1} + \xi_T)\Big],$$

$$\boldsymbol{d}_1(\boldsymbol{Q}_{\tilde{h}}) = \big[(-\phi\xi_2), (-\phi\xi_3), \ldots, (-\phi\xi_{T-1})\big], \text{ and}$$

$$\boldsymbol{\ell}_{\tilde{h}} = \boldsymbol{\Sigma}_v^{-1}\big(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{\mu}}\big)$$
$$= \Big[\frac{\tilde{y}_1 - m_{s_1} - \mu}{v_{s_1}}, \frac{\tilde{y}_2 - m_{s_2} - (1-\phi)\mu}{v_{s_2}}, \ldots, \frac{\tilde{y}_T - m_{s_T} - (1-\phi)\mu}{v_{s_T}}\Big]'.$$

Drawing from this posterior distribution is straightforward and efficient, using band back-substitution described in Kastner and Frühwirth-Schnatter (2014): (1) compute the Cholesky decomposition $\boldsymbol{Q}_{\tilde{h}} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is lower triangle; (2) solve $\boldsymbol{L}\boldsymbol{a} = \boldsymbol{\ell}_{\tilde{h}}$ for $\boldsymbol{a}$; and (3) solve $\boldsymbol{L}'\tilde{\boldsymbol{h}} = \boldsymbol{a} + \boldsymbol{e}$ for $\tilde{\boldsymbol{h}}$, where $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}_T)$.

Conditional on the log-volatilities $\{h_t\}$, we sample the AR(1) evolution parameters: the log-innovation precisions $\{\xi_t\}$, the autoregressive coefficient ϕ, and the unconditional mean μ. The precisions are distributed $[\xi_t|\eta_t] \sim \text{PG}(1, \eta_t)$ for $\eta_t = h_{t+1} - \mu - \phi(h_t - \mu)$, which we sample using the `rpg()` function in the `R` package `BayesLogit` (Polson et al., 2013). The Pólya-Gamma sampler is efficient: using only exponential and inverse-Gaussian draws, Polson et al. (2013) construct an accept-reject sampler for which the probability of acceptance is uniformly bounded below at 0.99919, which does not require any tuning. Next, we assume the prior $[(\phi + 1)/2] \sim \text{Beta}(a_\phi, b_\phi)$, which restricts $|\phi| < 1$ for stationarity, and sample from the full conditional distribution of ϕ using the slice sampler of Neal (2003). We select $a_\phi = 10$ and $b_\phi = 2$, which places most of the mass for the density of ϕ in $(0, 1)$ with a prior mean of $2/3$ and a prior mode of $4/5$ to reflect the likely presence of persistent volatility clustering. The prior for the global scale parameter is $\tau \sim C^+(0, \sigma_\epsilon/\sqrt{T})$, which implies $\mu = \log(\tau^2)$ is $[\mu|\sigma_\epsilon, \xi_\mu] \sim N(\log(\sigma_\epsilon^2/T), \xi_\mu^{-1})$ with $\xi_\mu \sim \text{PG}(1, 0)$. Including the initialization $h_1 \sim N(\mu, \xi_0^{-1})$ with $\xi_0 \sim \text{PG}(1, 0)$, the posterior dis-

tribution for μ is $\mu \sim N(Q_\mu^{-1}\ell_\mu, Q_\mu^{-1})$ with $Q_\mu = \xi_\mu + \xi_0 + (1-\phi)^2 \sum_{t=1}^{T-1} \xi_t$ and $\ell_\mu = \xi_\mu \log(\sigma_\epsilon^2/T) + \xi_0 h_1 + (1-\phi) \sum_{t=1}^{T-1} \xi_t (h_{t+1} - \phi h_t)$. Sampling ξ_μ and ξ_0 follows the Pólya-Gamma sampling scheme above.

Finally, we sample the discrete mixture component indicators s_t. The discrete mixture probabilities are straightforward to compute: the prior mixture probabilities are the mixing proportions given by Omori et al. (2007) and the likelihood is $\tilde{y}_t \overset{indep}{\sim} N(h_t + m_{s_t}, v_{s_t})$; see Kastner and Frühwirth-Schnatter (2014) for details.

3.6 Conclusions

Dynamic shrinkage processes provide a computationally convenient and widely applicable mechanism for incorporating adaptive shrinkage and regularization into existing models. By extending a broad class of global-local shrinkage priors to the dynamic setting, the resulting processes inherit the desirable shrinkage behavior, but with greater time-localization. The success of dynamic shrinkage processes suggests that other priors may benefit from log-scale or other appropriate representations, with or without additional dependence modeling.

As demonstrated in Sections 3.3 and 3.4, dynamic shrinkage processes are particularly appropriate for dynamic linear models, including trend filtering and time-varying parameter regression. In both settings, the dynamic linear models with dynamic horseshoe innovations outperform all competitors in simulated data, and produce reasonable and interpretable results for real data applications. Dynamic shrinkage processes may be useful in other dynamic linear

models, such as incorporating seasonality or change points with appropriately-defined (dynamic) shrinkage. Given the exceptional curve-fitting capabilities of the Bayesian trend filtering model (3.8) with dynamic horseshoe innovations (BTF-DHS), a natural extension would be to incorporate the BTF-DHS into more general additive, functional, or longitudinal data models in order to capture irregular or local curve features.

An important extension of the dynamic fundamental factor model of Section 3.4.2 is to incorporate a large number of assets, possibly with residual correlation among stock returns beyond the common factors of FF-5. Building upon Carvalho et al. (2011), a reasonable approach may be to combine a set of known factors, such as the Fama-French factors, with a set of unknown factors to be estimated from the data, where *both* sets of factor loadings are endowed with dynamic shrinkage processes to provide greater adaptability yet sufficient capability for shrinkage of irrelevant factors.

CHAPTER 4

FUNCTIONAL AUTOREGRESSION FOR SPARSELY SAMPLED DATA

Portions of this chapter were published in Kowal et al. (2017).

4.1 Introduction

We develop a hierarchical Gaussian process model for forecasting and inference of functional time series data. A *functional time series* is a time-ordered sequence of random functions, $Y_1, \ldots, Y_T$, on some compact index set $\mathcal{T} \subset \mathbb{R}^D$, typically with $D = 1$. Unlike existing methods, our approach is especially suited for sparsely or irregularly sampled curves, in which the functions $Y_t(\tau)$ are observed at a small number of possibly unequally-spaced points $\tau \in \mathcal{T}$, and for curves sampled with non-negligible measurement error, which occur frequently in financial applications. Applications of functional time series are abundant, including: daily or weekly interest rate curves as a function of time to maturity, such as daily Eurodollar futures contracts (Kargin and Onatski, 2008) and weekly yield curves (Hays et al., 2012; Kowal et al., 2016); yearly sea surface temperature as a function of time-of-year (Besse et al., 2000); yearly mortality and fertility rates as a function of age (Hyndman and Ullah, 2007); daily pollution curves as a function of time-of-day (Damon and Guillas, 2002; Aue et al., 2015); and a vast collection of spatio-temporal applications in which a time-dependent variable is measured as a function of spatial location (e.g., Cressie and Wikle, 2011). The primary goal of functional time series analysis is usually forecasting $\{Y_t\}$, but we are also interested in performing inference and obtaining an interpretable representation of the time evolution of $\{Y_t\}$.

The most prevalent model for functional time series data is the *functional autoregressive model of order 1*, written FAR(1):

$$Y_t - \mu = \Psi(Y_{t-1} - \mu) + \epsilon_t, \tag{4.1}$$

where $Y_t \in L^2(\mathcal{T})$, Ψ is a bounded linear operator on $L^2(\mathcal{T})$, $\epsilon_t \in L^2(\mathcal{T})$ is a sequence of independent mean zero random innovation functions with $\mathbb{E}||\epsilon_t||^2 < \infty$, and μ is the mean of $\{Y_t\}$ under stationarity. The FAR(1) model, developed by Bosq (2000), is an extension of two highly successful models: the functional linear model for function-on-function regression and the vector autoregressive model for multivariate time series, and has been successfully applied in a variety of applications. Importantly, the FAR(1) model provides a mechanism for modeling the evolution of $\{Y_t\}$ jointly over the entirety of the domain $\mathcal{T}$. More generally, (4.1) can be extended for multiple lags to the FAR(p) model: $Y_t - \mu = \sum_{\ell=1}^{p} \Psi_\ell(Y_{t-\ell} - \mu) + \epsilon_t$.

Existing approaches for estimating the FAR(p) model typically use an eigendecomposition of the empirical (contemporaneous and lagged) covariance operators (Damon and Guillas, 2002, 2005; Horváth and Kokoszka, 2012; Kokoszka, 2012) or kernel-based procedures for modeling the conditional expectation (Besse et al., 2000). A related approach is to estimate a multivariate time series model for the functional principal component (FPC) scores of the observed data (Aue et al., 2015). Extensions of the FAR(1) model for nonstationary functional time series are available, such as the time-dependent FAR kernels proposed in Chen and Li (2015).

In general, existing methods for FAR(p) are designed for functional data observed on dense grids without measurement error, and typically require presmoothing discretized functional observations. However, such procedures may

exhibit erratic behavior for sparse designs and are inappropriate in such settings. More generally, under an FAR(p) model that includes measurement error and discretization of the functional observations, we prove that the two most common approaches for functional data analysis—estimators that are linear in the FPC scores or the pre-smoothed observations—produce predictions that are inadmissible (in a decision theory sense). Indeed, the presence of measurement error fundamentally alters the behavior of the observable process: if an FAR process is observed with measurement error, then the observable process is no longer an FAR process, but rather a functional autoregressive moving average process (see Proposition 4.1). Even under dense designs, existing methods produce poor estimates of the FAR operator Ψ (Didericksen et al., 2012), which inhibits interpretability of the time evolution of $\{Y_t\}$, and do not provide finite-sample inference. We propose new methodology that simultaneously addresses all of these challenges.

We propose a general two-level hierarchy for modeling functional time series: an *observation equation* addresses measurement error and discretization of the functional data, while an *evolution equation* defines a process model for the underlying functional time series. The latent process is dynamically modeled as an FAR(p). We parsimoniously specify the FAR model with mean zero Gaussian process innovations, which are fully specified by covariance functions without parameterizing sample paths. The dynamic innovation process is further specified by a dynamic functional factor model. In contrast with standard approaches for Gaussian processes, this avoids selecting and estimating a parametric covariance function, and allows greater computational stability and efficiency, and broader applicability. Interpolating curves at unsampled locations and forecasting future curves are primary objectives in functional time series

modeling; the proposed model produces optimal (best linear) predictions under both sparse and dense designs in the presence of measurement error, even with the Gaussian assumption relaxed. We propose an efficient Gibbs sampling algorithm for estimation, inference, and forecasting. Extensive simulations demonstrate substantial improvements in forecasting performance and recovery of the autoregressive surface over competing methods, especially under sparse designs.

We apply our methodology to model and forecast *nominal* and *real* yield curves using daily U.S. data. For a given currency and level of risk of a debt, the nominal yield curve, $Y_t^N(\tau)$, describes the interest rate at time t as a function of the length of the borrowing period, or time to maturity, τ. Similarly, the real yield curve, $Y_t^R(\tau)$, corresponds to an interest rate that is adjusted for inflation. Both Y_t^N and Y_t^R may be modeled as functional time series. However, real yields are sparsely observed for each time t, and only at longer maturities, which is problematic for existing functional time series models. The proposed methods provide a natural hierarchical framework for modeling both nominal yield curves and real yield curves, and in both cases produce highly competitive forecasts.

Bayesian methods for functional time series are limited, with the exception of Laurini (2014) and Kowal et al. (2016). The primary contributions of this article are the following: (i) development of a hierarchical framework for FAR(p) (Section 4.2), which produces optimal (best linear) predictions under both sparse and dense designs in the presence of measurement error; (ii) a dynamic functional factor model for the innovation covariance, which is nonparametric, computationally convenient, and offers useful generalizations to non-

Gaussian distributions (Section 4.3); (iii) a procedure for model averaging over the lag, p, within a hierarchical FAR(p) model (Section 4.4); (iv) comparisons of the proposed methods to existing methods for FAR(p) using theoretical results (Section 4.5), an extensive simulation study (Section 4.6), and a real data application (Section 4.7); (v) a comparative forecasting study of daily U.S. nominal and real yield curve data (Section 4.7); and (vi) an efficient Gibbs sampling algorithm, which uses common full conditional distributions and existing `R` software (Appendix C). Details of our Gibbs sampling algorithm and additional theoretical and simulation results are in Appendix C.

4.2 Hierarchical Gaussian Processes for FAR

Let $Y_1, \ldots, Y_T$ be a time-ordered sequence of random functions in $L^2(\mathcal{T})$, where $\mathcal{T} \subset \mathbb{R}^D$ is a compact index set. We focus on $D = 1$ with $\mathcal{T} = [0, 1]$, but the methods can be developed more generally. For interpretability and computational convenience, we restrict our attention to the integral operators defined by $\Psi_\ell(Y)(\tau) = \int \psi_\ell(\tau, u) Y(u)\, du$, so the FAR($p$) model is

$$Y_t(\tau) - \mu(\tau) = \sum_{\ell=1}^{p} \int \psi_\ell(\tau, u) \left\{ Y_{t-\ell}(u) - \mu(u) \right\} du + \epsilon_t(\tau) \quad \forall \tau \in \mathcal{T}. \tag{4.2}$$

Using integral operators, the FAR(p) model resembles the functional linear model, in which $(Y_t - \mu)$ is regressed on $(Y_{t-1} - \mu), \ldots, (Y_{t-p} - \mu)$. The functional linear model is widely popular in functional data analysis, and has been extensively studied (e.g., Cardot et al., 1999; Ramsay, 2006).

In practice, model (4.2) is incomplete: the functional observations $\{Y_t\}$ are not observed directly, but rather via discrete samples of each curve, and typically with measurement error. Suppose that we observe $y_{i,t} \in \mathbb{R}$ sampled with

noise $\nu_{i,t}$ from $Y_t \in L^2(\mathcal{T})$:

$$y_{i,t} = Y_t(\tau_{i,t}) + \nu_{i,t} \tag{4.3}$$

for $i = 1, \ldots, m_t$, where $\tau_{1,t}, \ldots, \tau_{m_t,t}$ are the observation points of Y_t and $\nu_{i,t}$ is a mean zero measurement error with finite variance. Typically for functional data, m_t will be large and $\mathcal{T}_t = \{\tau_{1,t}, \ldots, \tau_{m_t,t}\}$ will be dense in $\mathcal{T}$. However, for our procedures, we allow m_t to be small for some (or all) t, with observation points $\mathcal{T}_o \equiv \cup_t \mathcal{T}_t$ dense or sparse in $\mathcal{T}$. Combining (4.3) with (4.2) for $p = 1$ and defining $\mu_t \equiv Y_t - \mu$, we obtain the two-level hierarchical model

$$\begin{cases} y_{i,t} = \mu(\tau_{i,t}) + \mu_t(\tau_{i,t}) + \nu_{i,t}, & i = 1, \ldots, m_t, \\ \mu_t(\tau) = \int \psi(\tau, u)\mu_{t-1}(u)\, du + \epsilon_t(\tau), & \forall \tau \in \mathcal{T} \end{cases} \tag{4.4}$$

for $t = 2, \ldots, T$, where we assume that $\{\nu_{i,t}\}$ and $\{\epsilon_t\}$ are mutually independent sequences.

The measurement error is a nontrivial component of model (4.4), which we demonstrate in the following proposition:

Proposition 4.1. *Let $Y_t - \mu = \sum_{\ell=1}^{p} \Psi_\ell(Y_{t-\ell} - \mu) + \epsilon_t$, and suppose that we observe $y_t = Y_t + \nu_t$, where $\{\epsilon_t\}$ and $\{\nu_t\}$ are independent white noise processes. Then the observable process $\{y_t\}$ follows a* functional autoregressive moving average *(FARMA) process of order (p, p).*

We define a FARMA process and prove Proposition 4.1 in Section C.4.1 of Appendix C. The implication of Proposition 4.1 is that, if the true model for Y_t is FAR(p), yet Y_t is observed with error, then the FAR(p) model for the observables is inappropriate. As a result, estimation of Ψ_ℓ will be inefficient and forecasting will deteriorate, due to both increased estimation error of Ψ_ℓ and model misspecification. By comparison, the hierarchical model decomposes the

observed data into a functional (autoregressive) process and measurement error, and in doing so circumvents the model misspecification issues implied by Proposition 4.1.

We model the random functions μ, ψ, and $\{\epsilon_t\}$ as *Gaussian processes*: $\mu \sim \mathcal{GP}(0, K_\mu)$, $\psi \sim \mathcal{GP}(0, K_\psi)$, and $\epsilon_t \overset{indep}{\sim} \mathcal{GP}(0, K_\epsilon)$, where the notation $\mathcal{GP}(m, K)$ denotes a Gaussian process with mean function m and covariance function K. Gaussian processes have a long history in machine learning (Rasmussen and Williams, 2006) and spatial statistics (Cressie and Wikle, 2011), and have seen increased application in functional data analysis, especially for hierarchical modeling (Behseta et al., 2005; Kaufman and Sain, 2010; Shi and Choi, 2011; Earls and Hooker, 2014). The conditional distribution of $\mu_t = Y_t - \mu$ is $[\mu_t|\mu_{t-1}, \psi, K_\epsilon] \sim \mathcal{GP}(\int \psi(\cdot, u)\mu_{t-1}(u)\, du, K_\epsilon)$, which models the evolution of μ_t and serves as the prior distribution for the observation level of (4.4). Notably, the model only requires *conditionally* Gaussian processes, and therefore may accommodate more general distributional assumptions, such as scale-mixtures of Gaussian distributions and stochastic volatility. Moreover, the posterior expectations derived from the hierarchical Gaussian process model are best linear predictors, and therefore are optimal among linear predictors for interpolation and forecasting of Y_t, even for non-Gaussian distributions (see Section 4.5). We assume $\nu_{i,t} \overset{iid}{\sim} N(0, \sigma_\nu^2)$ for the measurement errors; priors for σ_ν^2 and the parameters associated with K_μ, K_ϵ, and K_ψ will be discussed later.

4.2.1 Dynamic Linear Models for FAR(p)

For practical implementation of model (4.4), we must select a finite set of evaluation points, $\mathcal{T}_e \equiv \{\tau_1, \ldots, \tau_M\} \subset \mathcal{T}$, at which we wish to estimate, forecast, or perform inference on the random functions, in particular $\mu_t = Y_t - \mu$. Naturally, we assume that $\mathcal{T}_t \subseteq \mathcal{T}_e$ for all t, but this assumption may be relaxed. Notably, $\mathcal{T}_e$ provides a convenient structure for forecasting and inference of $y_{i,t}$ and $Y_t(\tau_{i,t})$ at the observations points $\tau_{i,t} \in \mathcal{T}_t$, as well as interpolation of Y_t at any unobserved points, $\tau^* \in \mathcal{T}_e \setminus \mathcal{T}_o$. By definition, for any Gaussian process $x \sim \mathcal{GP}(m, K)$ defined on $\mathcal{T}$, we have $\boldsymbol{x} \sim N(\boldsymbol{m}, \boldsymbol{K})$, where $\boldsymbol{x} = (x(\tau_1), \ldots, x(\tau_M))'$, $\boldsymbol{m} = (m(\tau_1), \ldots, m(\tau_M))'$, and $\boldsymbol{K} = \{K(\tau_i, \tau_k)\}_{i,k=1}^M$. This result is particularly useful for constructing an estimation procedure and deriving the optimality results of Section 4.5.

By selecting M large and $\mathcal{T}_e$ dense in $\mathcal{T}$, we can accurately approximate the integral in (4.4) using quadrature methods:

$$\int \psi(\tau, u)\mu_{t-1}(u)\, du \approx (\psi(\tau, \tau_1), \ldots, \psi(\tau, \tau_M))\, \boldsymbol{Q}\boldsymbol{\mu}_{t-1}, \tag{4.5}$$

where $\boldsymbol{Q}$ is a known quadrature weight matrix and $\boldsymbol{\mu}_{t-1} = (\mu_{t-1}(\tau_1), \ldots, \mu_{t-1}(\tau_M))'$. The approximation in (4.5) is important for computational tractability in estimation of both μ_t and ψ. Practical implementations of functional data methods require discretization or finite approximations; the quadrature approximation in (4.5) is a natural approach, and does not impose restrictive assumptions on the functional forms of ψ and μ_{t-1}. In addition, our simulation analysis suggests that the quadrature approximation does not noticeably inhibit estimation or forecasting, especially relative to existing FAR methods. In practice, the trapezoidal rule for computing $\boldsymbol{Q}$ works well, and for simulated data $M = 20$ is sufficiently large. We include a sensitivity analysis in Appendix C to assess the

effects of M on the approximation error in (4.5), which supports this choice of M.

Assuming $\mathcal{T}_o \subseteq \mathcal{T}_e$, let $\boldsymbol{Z}_t$ be the $m_t \times M$ incidence matrix that identifies the observations points observed at time t, i.e., $(\tau_{1,t}, \ldots, \tau_{m_t,t})' = \boldsymbol{Z}_t(\tau_1, \ldots, \tau_M)'$. We can write the hierarchical model (4.4) as a *dynamic linear model* (DLM; West and Harrison, 1997) in $\boldsymbol{\mu}_t$:

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\mu} + \boldsymbol{Z}_t\boldsymbol{\mu}_t + \boldsymbol{\nu}_t, & [\boldsymbol{\nu}_t|\sigma_\nu^2] \overset{indep}{\sim} N\left(\boldsymbol{0}, \sigma_\nu^2 \boldsymbol{I}_{m_t}\right) \text{ for } t = 1, \ldots, T, \\ \boldsymbol{\mu}_t = \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\epsilon}_t, & [\boldsymbol{\epsilon}_t|\boldsymbol{K}_\epsilon] \overset{indep}{\sim} N\left(\boldsymbol{0}, \boldsymbol{K}_\epsilon\right) \text{ for } t = 2, \ldots, T, \\ \boldsymbol{\mu}_1 \sim N(\boldsymbol{0}, \boldsymbol{K}_\epsilon), \end{cases} \tag{4.6}$$

where $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{m_t,t})'$, $\boldsymbol{\mu} = (\mu(\tau_1), \ldots, \mu(\tau_M))'$, $\boldsymbol{\Psi} = \{\psi(\tau_i, \tau_k)\}_{i,k=1}^M$, and $\boldsymbol{K}_\epsilon = \{K_\epsilon(\tau_i, \tau_k)\}_{i,k=1}^M$. Model (4.6) can be extended for multiple lags to the FAR(p) model by replacing the second level with $\boldsymbol{\mu}_t = \sum_{\ell=1}^p \boldsymbol{\Psi}_\ell \boldsymbol{Q} \boldsymbol{\mu}_{t-\ell} + \boldsymbol{\epsilon}_t$ for $\boldsymbol{\Psi}_\ell = \{\psi_\ell(\tau_i, \tau_k)\}_{i,k=1}^M$. The DLM formulation of the FAR(p) is useful for MCMC sampling, since efficient samplers exist for the vector-valued state variables, $\{\boldsymbol{\mu}_t\}$ (e.g., Durbin and Koopman, 2002). The proposed Gibbs sampling algorithm for model (4.6) (see Appendix C) is a moderate extension of traditional DLM samplers, and iteratively samples the state vectors $\{\boldsymbol{\mu}_t\}$, the measurement error variance σ_ν^2, the innovation covariance $\boldsymbol{K}_\epsilon$, and the unknown evolution matrix $\boldsymbol{\Psi}$. The DLM also facilitates non-Bayesian parameter estimation and forecasting, such as an EM algorithm for the latent state variables $\{\boldsymbol{\mu}_t\}$ with the parameters $\{\sigma_\nu^2, \boldsymbol{K}_\epsilon, \boldsymbol{\Psi}\}$ (e.g., Cressie and Wikle, 2011).

The connection between the hierarchical FAR model (4.4) and the DLM (4.6) is further illuminated by considering the autocovariance properties of the respective models. Recalling $\mu_t(\tau) = Y_t(\tau) - \mu(\tau)$, let $C_\ell(\tau_1, \tau_2) = \mathbb{E}\left[\mu_t(\tau_1)\mu_{t-\ell}(\tau_2)\right]$ be the lag-ℓ autocovariance function of $\{Y_t\}$, which is time-invariant under sta-

tionarity of $\{Y_t\}$. Under model (4.4) and assuming stationarity of $\{Y_t\}$, the lag-1 autocovariance function is equivalently $C_1(\tau_1, \tau_2) = \mathbb{E}[\mu_t(\tau_1)\mu_{t-1}(\tau_2)] = \mathbb{E}\left[\left\{\int \psi(\tau_1, u)\mu_{t-1}(u)\, du + \epsilon_t(\tau_1)\right\}\mu_{t-1}(\tau_2)\right] = \int \psi(\tau_1, u)C_0(u, \tau_2)\, du$. For $\ell \geq 1$, we have the more general recursion $C_\ell(\tau_1, \tau_2) = \int \psi(\tau_1, u)C_{\ell-1}(u, \tau_2)\, du$, from which it is clear that each C_ℓ is completely determined by the pair (ψ, C_0). Now let $\boldsymbol{C}_\ell = \mathbb{E}\left[\boldsymbol{\mu}_t\boldsymbol{\mu}_{t-\ell}'\right]$ be the lag-ℓ autocovariance matrix for the vector-valued time series $\{\boldsymbol{\mu}_t\}$ in (4.6). Under stationarity of $\{\boldsymbol{\mu}_t\}$, the lag-1 autocovariance matrix of $\boldsymbol{\mu}_t$ is $\boldsymbol{C}_1 = \mathbb{E}\left[\boldsymbol{\mu}_t\boldsymbol{\mu}_{t-1}'\right] = \mathbb{E}\left[\{\boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\epsilon}_t\}\boldsymbol{\mu}_{t-1}'\right] = \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{C}_0$. Notably, the relationship $\boldsymbol{C}_1 = \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{C}_0$ is an approximation to the continuous version, $C_1(\tau_1, \tau_2) = \int \psi(\tau_1, u)C_0(u, \tau_2)\, du$, using the same quadrature approximation as in (4.5). More generally, the matrix recursion $\boldsymbol{C}_\ell = \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{C}_{\ell-1}$ is a quadrature-based approximation to the continuous recursion, $C_\ell(\tau_1, \tau_2) = \int \psi(\tau_1, u)C_{\ell-1}(u, \tau_2)\, du$ for $\ell \geq 1$. Therefore, the evolution matrix $\boldsymbol{\Psi}\boldsymbol{Q}$ in the DLM (4.6) induces a discrete approximation to the autocovariance structure in the hierarchical FAR model (4.4).

The evolution equation of (4.6) resembles a VAR(1) on $\boldsymbol{\mu}_t = (\mu_t(\tau_1), \ldots, \mu_t(\tau_M))'$, but differs from a standard VAR on $\boldsymbol{y}_t$ for a few critical reasons. First, fitting a VAR to $\boldsymbol{y}_t$ is only well-defined if both the dimension m_t and the observation points $\mathcal{T}_t$ are fixed over time. If this does not hold, then imputation is necessary. Our procedure imputes automatically and optimally using the conditional mean function and the conditional covariance function of the corresponding Gaussian process. Second, the components of $\boldsymbol{y}_t$ are likely highly correlated due to the functional nature of the observations. Strong collinearity in VARs can cause overfitting and adversely affect forecasting and inference. In our model, the kernel function ψ is regularized using a smoothness prior (see Section 4.4), which mitigates the adverse effects of collinearity on estimation of ψ.

The smoothness prior on ψ is a nonstandard regularization technique for VARs, but is appropriate in this setting. Finally, the quadrature matrix, $\boldsymbol{Q}$, is absorbed into the VAR coefficient matrix $\boldsymbol{\Psi Q}$, and reweights the vector $\boldsymbol{\mu}_{t-1}$ using information from the evaluation points $\mathcal{T}_e$. This reweighting incorporates not only the vector values $\boldsymbol{\mu}_t$, but also the information that the components of $\boldsymbol{\mu}_t$ correspond to ordered elements of $\mathcal{T}_e$, which need not be equally spaced. The simulations of Section 4.6 demonstrate the substantial improvements in forecasting of our procedure relative to a VAR on $\boldsymbol{y}_t$.

4.3 A Dynamic Functional Factor Model for the Innovation Process

The standard approach for Gaussian process models is to select a parametric covariance function that only depends on a few parameters, and then estimate those parameters using either fully Bayesian methods or empirical Bayes (Rasmussen and Williams, 2006). The choice of the covariance function determines the properties of the sample trajectories, such as smoothness and periodicity, but notably does *not* imply a parametric form for the sample trajectories. Indeed, the FAR(1) model (4.6) may be estimated using these standard approaches; we provide one implementation in Section 4.6.

However, there are substantial computational limitations that accompany standard parametric covariance functions. Even when the covariance function is known up to some parameters $\boldsymbol{\rho}$, in general we cannot directly sample from the full conditional posterior distribution for $\boldsymbol{\rho}$. As a result, posterior sampling for $\boldsymbol{\rho}$ can be inefficient. Gaussian processes also require computation of the

$M \times M$ innovation covariance matrix $\boldsymbol{K}_\epsilon$, which must be inverted—both for evaluating the conditional likelihood of $\boldsymbol{\rho}$ and for sampling $\{\boldsymbol{\mu}_t\}$ and ψ. Most common choices for parametric covariance functions do not offer any simplifying structure for computing this inverse, which may be computationally inefficient and unstable. In addition, extensions for time-dependent covariance functions or non-Gaussian distributions are not readily available, and further increase the difficulties with posterior sampling.

We propose a low-rank, fully nonparametric approach for modeling the innovation covariance function. Using the *functional dynamic linear model* (FDLM) of Kowal et al. (2016), we estimate the unknown covariance function using a functional factor model, which does not require specification of a parametric form for the covariance function. This method avoids the need for inversion of the full $M \times M$ covariance matrix, and is more computationally stable and efficient. The integration of the FDLM into (4.6) retains the fully Bayesian hierarchical structure, and permits joint inference for all parameters via an efficient MCMC sampling algorithm. A *functional* factor model is most appropriate because ϵ_t is a Gaussian process with covariance function K_ϵ, so K_ϵ must be well-defined on $\mathcal{T} \times \mathcal{T}$. Notably, the FDLM offers convenient generalizations for stochastic volatility models (Kim et al., 1998) and more robust models using scale-mixtures of Gaussian distributions (Fernandez and Steel, 2000).

The FDLM decomposes the innovations ϵ_t into *factor loading curves* (FLCs), $\phi_j \in L^2(\mathcal{T})$, and time-dependent *factors*, $e_{j,t} \in \mathbb{R}$, for $j = 1, \ldots, J_\epsilon$:

$$\epsilon_t(\tau) = \sum_{j=1}^{J_\epsilon} e_{j,t}\phi_j(\tau) + \eta_t(\tau) \quad \forall \tau \in \mathcal{T}, \tag{4.7}$$

where J_ϵ is the number of factors and $\{\eta_t\}$ is the mean zero approximation error with $\eta_t \overset{iid}{\sim} \mathcal{GP}(0, K_\eta)$, where $K_\eta(\tau, u) = \sigma_\eta^2 \mathbf{1}(\tau = u)$ and $\mathbf{1}(\cdot)$ is the indicator

function. We model each FLC ϕ_j as a smooth function admitting the basis expansion $\phi_j(\tau) = \boldsymbol{b}_\phi'(\tau)\boldsymbol{\xi}_j$, where $\boldsymbol{b}_\phi$ is a J_ϕ-dimensional vector of known basis functions and $\boldsymbol{\xi}_j$ is an unknown vector of coefficients. For superior MCMC performance, we prefer the low-rank thin plate spline basis for $\boldsymbol{b}_\phi$ (e.g., Crainiceanu et al., 2005) with knot locations selected using the quantiles of the observation points, $\mathcal{T}_o$. We place a smoothness prior on each $\boldsymbol{\xi}_j$, which is expressed via a conditionally conjugate Gaussian distribution and is convenient for efficient posterior sampling (see the Appendix). The smoothness assumption typically produces more interpretable FLCs $\{\phi_j\}$ and can improve estimation for unobserved points $\tau^* \notin \mathcal{T}_o$. For the factors $\boldsymbol{e}_t = (e_{1,t}, \ldots, e_{J_\epsilon,t})'$, we assume $[\boldsymbol{e}_t|\boldsymbol{\Sigma}_e] \overset{indep}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$, with $\boldsymbol{\Sigma}_e = \text{diag}\left(\{\sigma_j^2\}_{j=1}^{J_\epsilon}\right)$ for simplicity. By comparison, the factors in Kowal et al. (2016) are time-dependent; we assume independence to obtain a special case of the FDLM in which the implied innovation process $\{\epsilon_t\}$ is an independent sequence, which also improves computational efficiency of the FDLM sampling algorithm. Importantly, we obtain a nonparametric, low-rank approximation to the innovation covariance, K_ϵ, with useful computational simplifications.

For identifiability, we order the factors according to variability of ϵ_t explained, $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_{J_\epsilon}^2 > 0$, and require orthonormality of the FLCs. It is computationally convenient to enforce the discrete orthonormality constraint $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_{J_\epsilon}$, where $\boldsymbol{\Phi} = \boldsymbol{B}_\phi\boldsymbol{\Xi}$ is the $M \times J_\epsilon$ matrix of FLCs evaluated at $\mathcal{T}_e$, $\boldsymbol{B}_\phi = (\boldsymbol{b}_\phi(\tau_1), \ldots, \boldsymbol{b}_\phi(\tau_M))'$ is the $M \times J_\phi$ matrix of basis functions evaluated at $\mathcal{T}_e$, and $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{J_\epsilon})$ is the $J_\phi \times J_\epsilon$ matrix of unknown FLC basis coefficients. The implied covariance matrix for $\boldsymbol{\epsilon}_t = (\epsilon_t(\tau_1), \ldots, \epsilon_t(\tau_M))'$ under (4.7) is $\boldsymbol{K}_\epsilon = \boldsymbol{\Phi}\boldsymbol{\Sigma}_e\boldsymbol{\Phi}' + \sigma_\eta^2\boldsymbol{I}_M$, conditional on $\{\phi_j, \sigma_j^2\}$ and σ_η^2. Importantly, the discretized orthonormality constraint offers a substantial simplification for computing the

inverse of $\boldsymbol{K}_\epsilon$ using the Woodbury identity:

$$\boldsymbol{K}_\epsilon^{-1} = \sigma_\eta^{-2}\boldsymbol{I}_M - \sigma_\eta^{-2}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}', \tag{4.8}$$

where $\tilde{\boldsymbol{\Sigma}}_e = \sigma_\eta^{-2}\left(\boldsymbol{\Sigma}_e^{-1} + \sigma_\eta^{-2}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)^{-1} = \text{diag}\left(\{\sigma_j^2/(\sigma_\eta^2+\sigma_j^2)\}_{j=1}^{J_\epsilon}\right)$. As a result, $\boldsymbol{K}_\epsilon^{-1}$ may be computed without any matrix inversions. By comparison, parametric covariance functions not only fail to offer computational simplifications for $\boldsymbol{K}_\epsilon^{-1}$, but also require *additional* computations of $\boldsymbol{K}_\epsilon^{-1}$ in the estimation of the covariance function parameters, $\boldsymbol{\rho}$. The FDLM sampling algorithm for the factors $\{e_{j,t}\}$, the FLCs $\{\phi_j\}$, and the variances $\{\sigma_j^2\}$ and σ_η^2 is computationally inexpensive and MCMC efficient. Note that the approximation error is a non-trivial addition to model (4.7): η_t is necessary for nondegeneracy of $\boldsymbol{K}_\epsilon$, which is invertible only when $\sigma_\eta^2 > 0$. And while $\sigma_\eta^2 > 0$ implies that the innovations ϵ_t, and therefore μ_t, are not smooth, we find that in practice, the sample paths of ϵ_t and μ_t do appear smooth for sufficiently small σ_η^2. Generalizations to non-nugget approximation error variance functions $K_\eta(\tau,u) = \sigma_\eta^2(\tau)\mathbf{1}(\tau = u)$ for $\sigma_\eta^2\colon \mathcal{T} \to \mathbb{R}^+$ are available, but may introduce additional model complexity and computational costs.

An important application of the FDLM simplification in (4.8) is given in Theorem 4.2, in which we derive a computationally convenient form for estimating the out-of-sample posterior distribution $[\mu_t(\tau^*)|\{\boldsymbol{y}_r\}_{r=1}^s]$ for $\tau^* \notin \mathcal{T}_e$, which includes as special cases the forecasting distribution ($s < t$), the filtering distribution ($s = t$), and the smoothing distribution ($s > t$).

4.4 Modeling the FAR Kernel

An accurate predictor of ψ is important not only for forecasting and inference, but also for interpreting the time evolution of $\{Y_t\}$. The likelihood for ψ is specified by the evolution equation in model (4.6), which may be extended for multiple lags. We select a Gaussian process prior for ψ, which encourages smoothness of the surface and produces more interpretable results. Using the basis approximation $\psi_\ell(\tau, u) = \boldsymbol{b}_0'(\tau, u)\boldsymbol{\theta}_{\psi_\ell}$, we place a Gaussian prior on $\boldsymbol{\theta}_{\psi_\ell}$, which induces a Gaussian process prior for ψ_ℓ. A tensor product basis $\boldsymbol{b}_0'(\tau, u) = (\boldsymbol{b}_\psi'(u) \otimes \boldsymbol{b}_\psi'(\tau))$ for $\boldsymbol{b}_\psi$ a J_ψ-dimensional vector of B-spline basis functions is computationally efficient in our setting, especially for large M. The details are presented in the Appendix. Since $J_\psi < M$, the evolution matrix $\boldsymbol{\Psi Q}$ in (4.6) has $J_\psi^2 < M^2$ unknown parameters, so the evolution equation in the DLM (4.6) has fewer parameters than a standard VAR(1) on $\boldsymbol{\mu}_t$. Notably, the posterior distribution for ψ_ℓ depends on $\boldsymbol{K}_\epsilon^{-1}$, which is computationally unstable for many common parametric covariance functions. By comparison, the nonparametric FDLM estimate of $\boldsymbol{K}_\epsilon^{-1}$ in (4.8) is computationally stable, which further stabilizes estimates of ψ_ℓ.

An important choice in the FAR(p) model is the maximum lag, p: a poor choice of p can produce suboptimal forecasts and reduce MCMC efficiency. A reasonable approach is to compare the DIC or marginal likelihoods for different choices of p. However, this requires recomputing the model for each choice of p, which can be computationally intensive. Similarly, Kokoszka and Reimherr (2013) propose a multistage hypothesis testing procedure based on asymptotic approximations and an FPC decomposition, but would require modification for the hierarchical Bayesian implementation of (4.6).

Our approach is to select a maximum lag under consideration, p_{max}, and assign each lag ℓ a state variable, $s_\ell \in \{0, 1\}$, for $\ell = 1, \ldots, p_{max}$, to assess whether or not ψ_ℓ is included in the model:

$$\mu_t(\tau) = \sum_{\ell=1}^{p_{max}} s_\ell \int \psi_\ell(\tau, u)\mu_{t-\ell}(u)\, du + \epsilon_t(\tau), \tag{4.9}$$

which extends Kuo and Mallick (1998) and Korobilis (2013b) to the FAR(p) setting. By averaging over the states $\{s_\ell\}_{\ell=1}^{p_{max}}$, the forecasts of model (4.9) are the model-averaged forecasts over the FAR(ℓ) models for $\ell = 1, \ldots, p_{max}$. Since we restrict $s_\ell \in \{0, 1\}$, rather than strongly shrinking ψ_ℓ toward zero, we can substantially improve computational efficiency: at each MCMC iteration, we sample $\{\boldsymbol{\mu}_t\}$ jointly from the FAR(p^*) extension of the DLM (4.6), where $p^* = \min\{\ell : s_{\ell+1} = \cdots = s_{p_{max}} = 0\}$ is the largest lag of nonzero autocorrelation.

The joint distribution of the states is $[s_1, s_2, \ldots, s_{p_{max}}] = [s_1] \prod_{\ell=2}^{p_{max}} [s_\ell | s_{\ell-1}, \ldots, s_1]$, where $[s_\ell | s_{\ell-1}, \ldots, s_1]$ is the probability that the lag ℓ autocorrelation term is included in the model, given whether the autocorrelation terms of the smaller lags $\ell - 1, \ldots, 1$ are included in the model. We assume that $s_\ell = 0$ implies that s_k is likely also zero for all $k > \ell$, which induces a more parsimonious model. In particular, we use the computationally convenient Markov assumption $[s_\ell | s_{\ell-1}, \ldots, s_1] = [s_\ell | s_{\ell-1}]$ with a small transition probability for $\mathbb{P}(s_\ell = 1 | s_{\ell-1} = 0) = q_{01}$. The reverse transition probability, $\mathbb{P}(s_\ell = 0 | s_{\ell-1} = 1) = q_{10}$, encourages smaller models when it is large. By default, we select $q_{01} = 0.01$, $q_{10} = 0.75$, and complete the joint prior distribution of $\{s_\ell\}_{\ell=1}^{p_{max}}$ with $\mathbb{P}(s_1 = 1) = 0.9$; for simulated data, the posterior does not appear to be sensitive to these choices.

4.5 Finite-Dimensional Optimality

The Gaussian assumptions in model (4.6) provide convenient posterior distributions for MCMC sampling and a useful framework for inference, but are not necessary for model (4.2). Suppose we relax the Gaussian assumption to $\epsilon_t \sim \mathcal{SP}(0, K_\epsilon)$, where $\mathcal{SP}(m, K)$ denotes a second-order stochastic process with mean function m and covariance function K. Similarly, let $\nu_{i,t}$ be a mean zero random variable with variance σ_ν^2 and let $\mu_1 \equiv \epsilon_1$. Given a finite set of evaluation points, $\mathcal{T}_e \subset \mathcal{T}$, model (4.4) implies the distribution-free DLM

$$\begin{cases} \boldsymbol{y}_t = \boldsymbol{Z}_t\boldsymbol{\mu} + \boldsymbol{Z}_t\boldsymbol{\mu}_t + \boldsymbol{\nu}_t, & \mathbb{E}[\boldsymbol{\nu}_t|\sigma_\nu^2] = 0,\ \mathrm{Cov}[\boldsymbol{\nu}_t|\sigma_\nu^2] = \sigma_\nu^2\boldsymbol{I}_{m_t}, \\ \boldsymbol{\mu}_t = \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\epsilon}_t, & \mathbb{E}[\boldsymbol{\epsilon}_t|\boldsymbol{K}_\epsilon] = 0,\ \mathrm{Cov}[\boldsymbol{\epsilon}_t|\boldsymbol{K}_\epsilon] = \boldsymbol{K}_\epsilon, \end{cases} \tag{4.10}$$

under the integral approximation (4.5), where the vectors and matrices are defined as before and $\boldsymbol{\mu}_1 \equiv \boldsymbol{\epsilon}_1$. Since this holds for any finite set of evaluation points $\mathcal{T}_e \subset \mathcal{T}$, we may consider the DLM (4.10) to be a *collection* of models indexed by the evaluation points, $\mathcal{T}_e$. The error sequences, $\{\boldsymbol{\nu}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$, are assumed to be uncorrelated, rather than independent. If we additionally assume Gaussianity of $\{\boldsymbol{\nu}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$, then the uncorrelatedness implies independence, and model (4.10) becomes model (4.6). Extensions for the FAR(p) models are similar. The results below also hold for time-dependent variances for $\boldsymbol{\nu}_t$ and $\boldsymbol{\epsilon}_t$.

Let d be an estimator of $\delta \in L^2(\mathcal{T})$, and consider the squared error loss using the Euclidean norm: $\mathcal{L}_e(\delta, d) = (\boldsymbol{\delta} - \boldsymbol{d})'(\boldsymbol{\delta} - \boldsymbol{d})$, where $\mathcal{L}_e$ is indexed by the set of evaluation points, $\mathcal{T}_e$, at which δ and d are evaluated to form the corresponding vectors $\boldsymbol{\delta}$ and $\boldsymbol{d}$. When $\mathcal{T}_e$ is an equally-spaced fine grid on $\mathcal{T}$, the loss function $\mathcal{L}_e$ will approximate the usual loss function for functional data, $\mathcal{L}_{L^2}(\delta, d) = \int(\delta(u) - d(u))^2\,du$, for most reasonable choices of δ and d (up to a rescaling by $M = |\mathcal{T}_e|$). In a standard Bayesian analysis, the goal would be

to minimize the posterior risk, $\mathbb{E}[\mathcal{L}_e(\delta, d)|\{\boldsymbol{y}_t\}]$, for which the solution is the posterior expectation, $d = \mathbb{E}[\delta|\{\boldsymbol{y}_t\}]$. Indeed, the estimators discussed below minimize the posterior risk under the Gaussian assumptions of model (4.6). However, by relaxing the distributional assumptions in (4.10) to increase the generality of the model, we no longer have sufficient information to compute posterior distributions or posterior moments. In addition, it is difficult to compare Bayesian and non-Bayesian procedures under the posterior risk, and most procedures for functional time series modeling are non-Bayesian. Therefore, we consider the overall risk $\mathcal{R}_e(\delta, d) = \mathbb{E}[\mathcal{L}_e(\delta, d)]$, which is the expected value of the posterior risk with respect to the sampling distribution. As with the loss function $\mathcal{L}_e$, the risk function $\mathcal{R}_e$ is indexed by the evaluation points, $\mathcal{T}_e$; we seek to minimize $\mathcal{R}_e$ for any choice of $\mathcal{T}_e$.

Let $\mathcal{D}_t = \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1\} \cup \mathcal{D}_0$ be the information available at time t, where $\mathcal{D}_0$ represents the information prior to time $t = 1$.

Theorem 4.1. *For* any *finite set of evaluation points $\mathcal{T}_e \subset \mathcal{T}$, the* unique best linear predictor *of the conditional random vector $\boldsymbol{\delta} \sim [\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$, where $\boldsymbol{\delta}, \boldsymbol{Y} \subseteq \mathcal{D}_T \cup \{\mu_t(\tau) : \tau \in \mathcal{T}_e, t = 1, \ldots, T\}$ and $\boldsymbol{\Theta} = \{\mu, \sigma_\nu^2, \psi, K_\epsilon\}$, under the risk $\mathcal{R}_e$ and conditional on model* (4.4) *with the integral approximation* (4.5)*, is the conditional expectation $\hat{\boldsymbol{\delta}}(\boldsymbol{Y}|\boldsymbol{\Theta}) \equiv \mathbb{E}[\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$ as computed under model* (4.6)*.*

The proof of Theorem 4.1 is in the Appendix, and extends fundamental results for vector-valued DLMs. The best linear predictors of Theorem 4.1 equivalently minimize the risk $\mathcal{R}(\delta, d) = \sup_{\mathcal{T}_e} \mathcal{R}_e(\delta, d)$ among all linear estimators, where the $\sup$ is taken over all finite $\mathcal{T}_e \subset \mathcal{T}$. The most useful examples of $[\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$ in Theorem 4.1 are the forecasting distributions $[\boldsymbol{y}_{t+h}|\mathcal{D}_t, \boldsymbol{\Theta}]$ and $[\boldsymbol{\mu}_{t+h}|\mathcal{D}_t, \boldsymbol{\Theta}]$ for $h > 0$, the smoothing distributions $[\boldsymbol{\mu}_t|\mathcal{D}_T, \boldsymbol{\Theta}]$, and the filter-

ing distributions $[\boldsymbol{\mu}_t|\mathcal{D}_t, \boldsymbol{\Theta}]$, for $t = 1, \ldots, T$. Theorem 4.1 depends on the observation points $\mathcal{T}_o$ only via the assumption that $\boldsymbol{Z}_t$ is known. In general, we assume $\mathcal{T}_o \subseteq \mathcal{T}_e$, so $\boldsymbol{Z}_t$ is an incidence matrix and therefore known. Theorem 4.1 does *not* require $\mathcal{T}_o$ to become arbitrarily dense in $\mathcal{T}$, and is valid for both sparse and dense designs. For implementation, we compute the relevant expectations within the Gibbs sampling algorithm (see Appendix C), and then average over the Gibbs sample of $\boldsymbol{\Theta}$. Alternatively, an EM algorithm could be used to estimate the relevant expectations (Cressie and Wikle, 2011).

There is no intrinsic reason to restrict the estimators to linearity. However, several popular competing methods are linear, and therefore are dominated by the conditional expectations computed from model (4.6) whenever the estimators are distinct. More formally:

Corollary. *Consider a basis expansion of the observations $\boldsymbol{y}_t \approx \boldsymbol{B}_t\boldsymbol{\theta}_t$, where $\boldsymbol{B}_t' = (\boldsymbol{b}(\tau_{1,t}), \ldots, \boldsymbol{b}(\tau_{m_t,t}))$, $\boldsymbol{b}$ is a known J-dimensional vector of basis functions, and $\boldsymbol{\theta}_t$ is the corresponding J-dimensional vector of unknown basis coefficients. If the estimator $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}_t$ is linear in $\boldsymbol{y}_t$, then estimates or forecasts of the form $\boldsymbol{H}\hat{\boldsymbol{\theta}}_t + \boldsymbol{h}$, conditional on the matrix $\boldsymbol{H}$ and the vector $\boldsymbol{h}$, are inadmissible for all $[\boldsymbol{\delta}|\boldsymbol{Y}]$ whenever $\boldsymbol{H}\hat{\boldsymbol{\theta}}_t + \boldsymbol{h} \neq \hat{\boldsymbol{\delta}}(\boldsymbol{Y}|\boldsymbol{\Theta})$.*

The most important application of Corollary 4.5 is to characterize the inadmissibility of procedures based on FPC scores. In the notation of Corollary 4.5, let $\boldsymbol{b}$ be the FPC basis, which we assume is fixed and known. The components of $\boldsymbol{\theta}_t$ correspond to the FPC scores, defined by $\theta_{j,t} = \int \{Y_t(u) - \mu(u)\} b_j(u)\, du = \int \mu_t(u) b_j(u)\, du$. There are two standard approaches for computing FPC scores: quadrature methods for dense designs absent measurement error, and the PACE procedure of Yao et al. (2005), which uses conditional expectations under a

Gaussian assumption and applies more generally. In both cases, the FPC scores are linear in $\boldsymbol{y}_t$, so Corollary 4.5 applies.

Among functional time series methods, the most pertinent procedures are Aue et al. (2015) and Hyndman and Ullah (2007). Aue et al. (2015) provide the more general framework, in which they compute the best linear predictors for the FPC scores, and then forecast the FPC scores using multivariate time series methods. For time series methods that are *linear* in the FPC scores, such forecasts are inadmissible. While Aue et al. (2015) undoubtedly provide a simple yet general framework for forecasting a functional time series, the simulations of Section 4.6 confirm the consequence of inadmissibility on forecasting performance.

Corollary. *Consider the common functional data pre-processing procedure in which the discrete, noisy observations, $\boldsymbol{y}_t$, are replaced by estimated functions evaluated on a fine grid, $\hat{\boldsymbol{y}}_t$, and then estimates and forecasts are computed using the functional "data" $\hat{\boldsymbol{y}}_t$. If $\hat{\boldsymbol{y}}_t$ is linear in $\{\boldsymbol{y}_t\}$, then any estimator or forecast linear in $\{\hat{\boldsymbol{y}}_t\}$ is inadmissible for all $[\boldsymbol{\delta}|\boldsymbol{Y}]$ whenever $\hat{\boldsymbol{y}}_t \neq \hat{\boldsymbol{\delta}}(\boldsymbol{Y}|\boldsymbol{\Theta})$.*

Typically, $\hat{\boldsymbol{y}}_t$ is estimated using splines or kernel smoothers, both of which are linear in $\boldsymbol{y}_t$. As an application of Corollary 4.5, the simple forecasting method of fitting a VAR to $\hat{\boldsymbol{y}}_t$ evaluated on a grid of points, conditional on the VAR coefficient matrix, is inadmissible.

Corollary. *The unique best linear predictor of $[\mu_t(\tau^*)|\mathcal{D}_s]$ for any times t, s and any point $\tau^* \in \mathcal{T}$ is the corresponding expectation under model* (4.6).

Model (4.6) achieves the optimality of a kriging estimator for interpolation of any point $\tau^* \in \mathcal{T}$, simply by adding τ^* to the evaluation set $\mathcal{T}_e$.

In practice, we need not include all such τ^* in $\mathcal{T}_e$: we can estimate the out-of-sample posterior distribution $[\mu_t(\tau^*)|\mathcal{D}_s]$ for $\tau^* \notin \mathcal{T}_e$ by sampling from the out-of-sample full conditional distribution $\left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}, \mathcal{D}_s\right]$ within the Gibbs sampler, and then averaging over the Gibbs sample of $\{\boldsymbol{\mu}_r\}_{r=1}^T$ and $\boldsymbol{\Theta}$. Let $\boldsymbol{\psi}'(\tau^*) \equiv (\psi(\tau^*, \tau_1), \ldots, \psi(\tau^*, \tau_M))$ and $\boldsymbol{\phi}'(\tau^*) \equiv (\phi_1(\tau^*), \ldots, \phi_{J_\epsilon}(\tau^*))$. In the special case of model (4.4) and using the FDLM (4.7), we have the following computationally efficient alternative for state space imputation:

Theorem 4.2. *Suppose $\tau^* \in \mathcal{T}$ such that $\tau^* \notin \mathcal{T}_e$. Under the FDLM* (4.7) *and conditional on model* (4.4) *with the integral approximation* (4.5)*, the out-of-sample full conditional distribution of $\mu_t(\tau^*)$ is $\left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}, \mathcal{D}_s\right] \sim N\left(m_t(\tau^*), K_t(\tau^*)\right)$, where $m_t(\tau^*) = \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\phi}'(\tau^*)\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'\left(\boldsymbol{\mu}_t - \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1}\right)$ and $K_t(\tau^*) = \sigma_\eta^2 + \sigma_\eta^2\boldsymbol{\phi}'(\tau^*)\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\phi}(\tau^*)$.*

The proof of Theorem 4.2 and extensions for $p > 1$ are in Appendix C. Using Theorem 4.2, we can efficiently estimate the out-of-sample posterior distribution $[\mu_t(\tau^*)|\mathcal{D}_s]$ with minimal adjustments to the Gibbs sampling algorithm (see Appendix C). Theorem 4.2 builds upon the approximation in (4.5) and the computational simplifications of the FDLM to produce simple and efficient moment calculations for the full conditional distributions without expanding the dimension of the state vector, M. Note that for implementation purposes, the terms $\boldsymbol{\mu}_t$ and $\boldsymbol{\mu}_{t-1}$ appearing in $m_t(\tau^*)$ are assumed to be sampled from the full conditional distribution $\left[\{\boldsymbol{\mu}_r\}_{r=1}^T|\boldsymbol{\Theta}, \mathcal{D}_s\right]$.

4.6 Simulations

We conducted extensive simulations to evaluate the proposed methods for FAR(p) relative to several competitive alternatives. We are particularly interested in one-step forecasting and recovery of the FAR kernel ψ_1, and in how the associated performance varies with the sample size T, the location and number of the observation points $\tau_{1,t}, \ldots, \tau_{m_t,t}$, the kernel ψ_1, and the smoothness of the innovation process ϵ_t. We also assess the performance of the model averaging procedure of Section 4.4 for $p \in \{1, 2\}$, and compare the nonparametric FDLM approach of Section 4.3 with a more standard parametric Gaussian process implementation.

4.6.1 Sampling Designs

For all simulations, the mean function is $\mu(\tau) = \frac{1}{10}\tau^3 \sin(2\pi\tau)$, which produces the dominate shape in the rightmost panels of Figure 4.1. The measurement errors are identically distributed for all simulations: $\nu_{i,t} \overset{iid}{\sim} N(0, \sigma_\nu^2)$ with $\sigma_\nu = 0.002$. We vary the sample size from small ($T = 50$) to large ($T = 350$) for the FAR(1) simulations, and use a moderate sample size ($T = 125$) for the FAR(2) simulation. The FAR(1) kernel used for Figure 4.1 is the *Bimodal-Gaussian* kernel, $\psi(\tau, u) \propto \frac{0.75}{\pi(0.3)(0.4)} \exp\{-(\tau-0.2)^2/(0.3)^2 - (u-0.3)^2/(0.4)^2\} + \frac{0.45}{\pi(0.3)(0.4)} \exp\{-(\tau-0.7)^2/(0.3)^2 - (u-0.8)^2/(0.4)^2\}$, following Wood (2003); see Appendix C for a plot of the Bimodal-Gaussian kernel. We also present results for the *Linear-τ* kernel, $\psi(\tau, u) \propto \tau$, and the *Linear-$u$* kernel, $\psi(\tau, u) \propto u$. Each kernel is rescaled according to a pre-specified squared norm, $C_{\psi_\ell} = \int\int \psi_\ell^2(\tau, u)\, d\tau\, du$, with $\sum_{\ell=1}^{p} C_{\psi_\ell} < 1$ for stationarity. We select $C_{\psi_1} = 0.8$ for the FAR(1) simu-

lations and use $(C_{\psi_1}, C_{\psi_2}) = (0.4, 0.2)$ for the FAR(2) simulation; smaller values of C_{ψ_ℓ} produce similar comparative results, but the forecasting performance deteriorates for all methods. For the innovation process, ϵ_t, we consider both smooth and non-smooth Gaussian processes. We use the covariance function parametrization $K_\epsilon = \sigma^2 R_\rho$, where R_ρ is the Matérn correlation function $R_\rho(\tau, u) = \{2^{\rho_1 - 1}\Gamma(\rho_1)\}^{-1} (||\tau - u||/\rho_2)^{\rho_1} K_{\rho_1}(||\tau - u||/\rho_2)$, $\Gamma(\cdot)$ is the gamma function, K_{ρ_1} is the modified Bessel function of order ρ_1, and $\boldsymbol{\rho} = (\rho_1, \rho_2)$ are parameters (Matérn, 2013). We let $\sigma = 0.01$ and $\boldsymbol{\rho} = (\rho_1, 0.1)$, with $\rho_1 = 2.5$ for smooth (twice-differentiable) sample paths and $\rho_1 = 0.5$ for non-smooth (continuous, non-differentiable) sample paths.

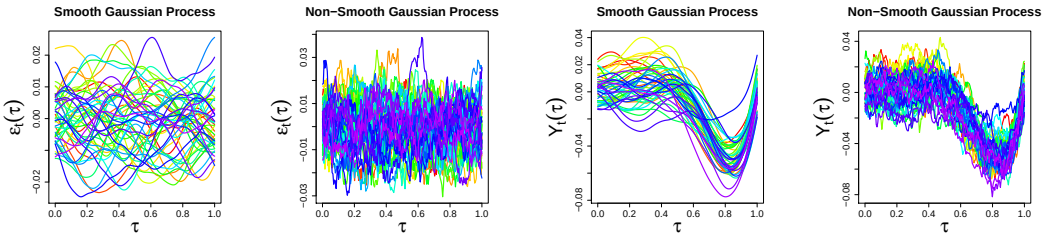


Figure 4.1: Sample paths of ϵ_t and $Y_t = \mu_t + \mu$ as a function of τ, where ϵ_t is a Gaussian process with the Matérn correlation function, $\boldsymbol{\rho} = (\rho_1, 0.1)$, $\sigma = 0.01$, and Y_t is generated using the Bimodal-Gaussian FAR(1) kernel, $t = 1, \ldots, T = 50$. The curves are time-ordered by color (from red/orange to blue/violet). **Left to right:** $\epsilon_t(\tau), \rho_1 = 2.5$; $\epsilon_t(\tau), \rho_1 = 0.5$; $Y_t(\tau), \rho_1 = 2.5$; $Y_t(\tau), \rho_1 = 0.5$. Note that we do not observe Y_t directly, but rather $y_{i,t} = Y_t(\tau_{i,t}) + \nu_{i,t}$, where $\nu_{i,t} \sim N(0, \sigma_\nu^2)$ is measurement error with $\sigma_\nu = \sigma/5 = 0.002$ and $\mathcal{T}_t = \{\tau_{1,t}, \ldots, \tau_{m_t,t}\}$ are the observation points at time t.

We consider three sampling designs for the observation points: *dense*, *sparse-random*, and *sparse-fixed*. In each case, the set of evaluation points, $\mathcal{T}_e$, is an equally-spaced grid of $M = 30$ points on $\mathcal{T} = [0, 1]$. The *dense* design uses $m_t = 25$ equally-spaced observation points on $[0, 1]$ for all t, for which the results are representative of denser ($m_t \gg 25$) designs and similar to those of Didericksen et al. (2012); see Appendix C. The *sparse-random* design is generated by

first sampling each m_t from a zero-truncated Poisson (5) distribution, and then sampling $\tau_{1,t}, \ldots, \tau_{m_t,t}$ without replacement from $\mathcal{T}_e$. This is a common design in sparse functional data, in which m_t may be small for some t, but $\mathcal{T}_o$ is dense in $\mathcal{T}$. The *sparse-fixed* design uses $m_t = 8$ equally-spaced points in $\mathcal{T}$. This is the most challenging design, and one for which multivariate time series methods should be most competitive with functional time series methods. Comparatively, the sparse settings are similar to the dense setting, but with additional missing observations.

4.6.2 Competing Estimators

Within the proposed framework and using the FDLM of Section 4.3 for the innovation covariance function, we compute forecasts for $p = 1$ (FDLM-FAR(1)), and in the FAR(2) simulation, for $p = 2$ (FDLM-FAR(2)) and $p = 3$ (FDLM-FAR(3)). We also compute forecasts using the model averaging procedure with $p_{max} = 4$ (FDLM-FAR(p)). To assess the performance of the FDLM implementation, we compute forecasts using model (4.6) with a parametric covariance function for $K_\epsilon = \sigma^2 R_\rho$ (GP-FAR(1)). We use the Matérn correlation function for R_ρ, with $\rho_1 = 2.5$ as in the smooth Gaussian process simulations, and use the priors $\sigma^{-2} \sim$ Gamma $(10^{-3}, 10^{-3})$ and $\rho_2 \sim$ Uniform $(0, U_{\rho_2})$, where U_{ρ_2} is the maximum value of ρ_2 for which the correlation function R_ρ is less than 0.99 for all pairs of evaluation points. These models are implemented using the Gibbs sampling algorithm provided in Appendix C, and estimates are based on 5,000 MCMC simulations after a burn-in of 5,000. For the large sample setting ($T = 350$), the mean computation time per 1,000 MCMC simulations was 2.3 minutes for FDLM-FAR(1) and 4.4 minutes for GP-FAR(1). The computing

times are calculated on a 64-bit Windows machine with a 2.40-GHz Intel core i7-4700MQ processor with 8 GB of RAM, and the code is written in `R`.

We consider several important competing methods. Let $\hat{\boldsymbol{y}}_{t+1}$ denote the one-step forecast at time t. For baseline comparisons, we use the random-walk (RW) forecast, $\hat{\boldsymbol{y}}_{t+1} = \boldsymbol{y}_t$, and the mean (Mean) forecast, $\hat{\boldsymbol{y}}_{t+1} = \hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}}$ is a smooth estimate of the mean of $\{\boldsymbol{y}_s\}_{s=1}^t$. We estimate $\hat{\boldsymbol{\mu}}$ using a B-spline basis expansion via the function `meanfd()` in the `R` package `fda` (Ramsay et al., 2014). Both estimators are robust against overfitting, and the mean forecast is optimal when $\psi = 0$. We also compute the one-step forecast based on a VAR(1) fit to $\{\boldsymbol{y}_s\}_{s=1}^t$ (VAR-Y). In the sparse-random design, the observations $\boldsymbol{y}_t$ were used to linear interpolate on $\mathcal{T}_e$ prior to fitting the VAR. In the sparse-fixed design, the VAR was fit to the observation points, and then forecasts for the evaluation points were computed by fitting a spline to the VAR forecasts of the observation points. For additional comparisons, we computed forecasts from a simple exponential smoother (SES) applied pointwise to each component of $\boldsymbol{y}_t$, i.e., each time series $\{y_{j,t}\}_{t=1}^T$. The SES forecasts are implemented using the `ses` function in the `R` package `forecast` (Hyndman and Khandakar, 2008), with an identical imputation scheme as VAR-Y. We also considered two functional data methods. First, we used the Estimated Kernel procedure outlined in Horváth and Kokoszka (2012), which estimates ψ_ℓ in (4.2) using FPCs (FAR Classic); we fix $p = 1$ for simplicity. This method has well-studied theoretical properties and is a useful baseline for FAR models. Second, we implemented the method of Aue et al. (2015), which we briefly described in Section 4.5, using a VAR(1) on the FPC scores (VAR-FPC). We compute the FPCs using the `fda` package in `R` with B-spline basis functions. To avoid the ill-conditioned estimators discussed in Horváth and Kokoszka (2012), we regularize via basis truncation, using 8 equally-spaced in-

terior knots. The number of components is selected to explain at least 95% of the variability in $\{\boldsymbol{y}_t\}$. For the sampling designs considered here, this approach works well. Finally, we report the oracle forecast (FAR Oracle) computed using the true one-step forecasts $\mathbb{E}\left[\mu_t(\tau)|\{\psi_\ell, \mu_{t-\ell}\}_{\ell=1}^p\right] = \sum_{\ell=1}^p \int \psi_\ell(\tau, u)\mu_{t-\ell}(u)\, du$ within the simulation, where $\{\psi_\ell\}_{\ell=1}^p$ are the FAR kernels from the simulation specification, $\{\mu_t\}$ are the simulated values of the latent FAR process, and the integral is approximated using the trapezoidal rule with $M = 200$ grid points. The oracle forecast is not actually an estimator, and is unaffected by sparsity or small sample sizes.

We estimate the one-step forecasts $[\boldsymbol{y}_{T+h}|\boldsymbol{y}_{1:(T+h-1)}]$, $h = 1, \ldots, 25$, for all estimators under consideration, and compare them using the mean squared forecast error $MSFE_e = \frac{1}{25M}\sum_{h=1}^{25} ||\boldsymbol{Y}_{T+h} - \hat{\boldsymbol{Y}}_{T+h}||^2$ where $\boldsymbol{Y}_{T+h} = (Y_{T+h}(\tau_1), \ldots, Y_{T+h}(\tau_M)))'$, which measures the one-step forecasting performance at the *evaluation* points, and the mean squared error $MSE_{\psi_1} = \frac{1}{M^2}\sum_{i=1}^M \sum_{k=1}^M \{\psi_1(\tau_i, \tau_k) - \hat{\psi}_1(\tau_i, \tau_k)\}^2$, which measures the recovery of the lag-1 kernel ψ_1. Because $\mathcal{T}_e$ is relatively dense in $\mathcal{T}$, $MSFE_e$ and MSE_{ψ_1} approximate the integrated squared errors $\int \{Y_{T+h}(u) - \hat{Y}_{T+h}(u)\}^2\, du$ and $\int\int \{\psi_1(\tau, u) - \hat{\psi}_1(\tau, u)\}^2\, d\tau\, du$, respectively. Estimators $\hat{\psi}_1$ are available only for the proposed methods and FAR Classic. For computational convenience in the proposed methods, we update $\{\boldsymbol{\mu}_t\}_{t=1}^{T+h-1}$ using all of the data $\boldsymbol{y}_{1:(T+h-1)}$, but sample all other parameters only conditional on $\boldsymbol{y}_{1:T}$. DLM updating algorithms provide recursive one-step forecasts for $\boldsymbol{\mu}_t$, but in general there are no convenient updating algorithms for the other parameters. In practice, this is not a problem, but suggests that our simulation analysis may underestimate the performance of the proposed model.

4.6.3 Results

We computed $MSFE_e$ and MSE_{ψ_1} under a variety of sampling designs, each for $N = 50$ simulations, and present the results for a few important cases in Figures 4.2 and 4.3, respectively. The figures are color-coded: multivariate methods are green, existing functional data methods are red, the proposed methods are blue, and the oracle is gold.

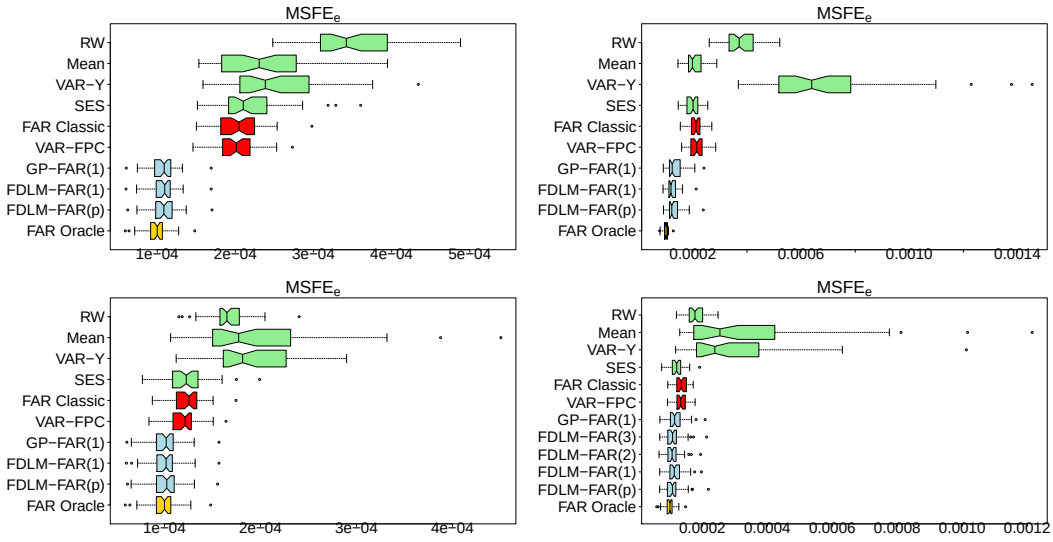


Figure 4.2: $MSFE_e$ under various designs. **Top left:** FAR(1), $T = 350$, sparse-random design with the Linear-u kernel and smooth GP innovations. **Top right:** FAR(1), $T = 50$, sparse-random design with the Bimodal-Gaussian kernel and non-smooth GP innovations. **Bottom left:** FAR(1), $T = 350$, sparse-fixed design with the Bimodal-Gaussian kernel and smooth GP innovations. **Bottom right:** FAR(2), $T = 125$, sparse-fixed design with Bimodal-Gaussian and Linear$-\tau$ kernels and smooth GP innovations. The proposed methods provide superior forecasts and nearly achieve the oracle performance, despite the presence of sparsity.

For the sparse designs in Figure 4.2, the proposed methods are all superior to the competitors, and in some cases nearly achieve the oracle performance, even though the oracle is unaffected by sparsity. Figure 4.3 shows that the proposed methods also offer a substantial improvement in ψ_1 estimation. Importantly, the

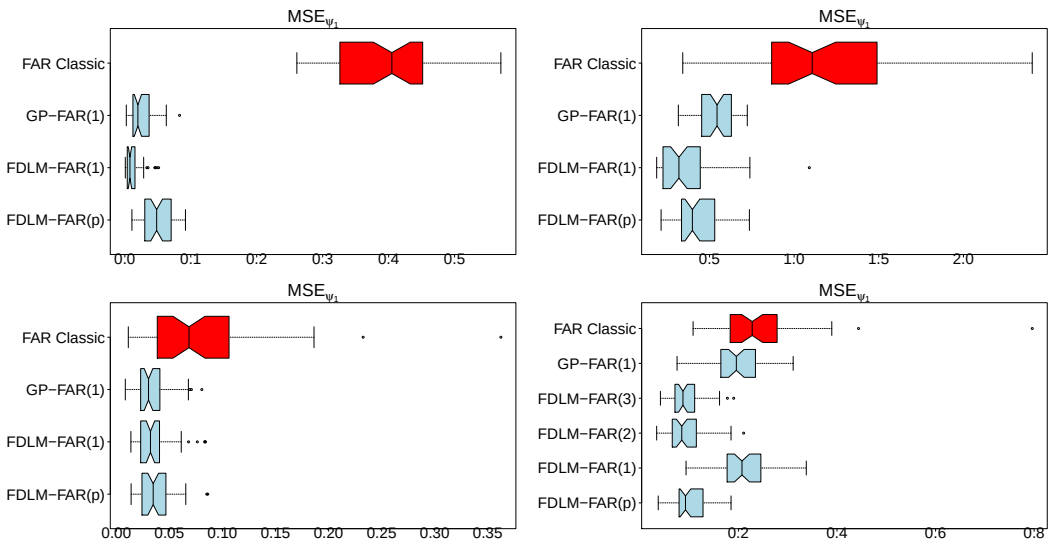


Figure 4.3: MSE_{ψ_1} under various designs. **Top left:** FAR(1), $T = 350$, sparse-random design with the Linear-u kernel and smooth GP innovations. **Top right:** FAR(1), $T = 50$, sparse-random design with the Bimodal-Gaussian kernel and non-smooth GP innovations. **Bottom left:** FAR(1), $T = 350$, sparse-fixed design with the Bimodal-Gaussian kernel and smooth GP innovations. **Bottom right:** FAR(2), $T = 125$, sparse-fixed design with Bimodal-Gaussian and Linear$-\tau$ kernels and smooth GP innovations. Estimates of ψ_1 are far superior for the proposed methods, including the FAR(p) with model averaging.

proposed model with model averaging is competitive with the known p model for both forecasting and estimation of ψ_1. The model averaging procedure of Section 4.4 typically identifies the true p with high probability, with a mild tendency to overestimate p. However, this behavior is encouraging: the bottom right panel of Figure 4.3, in which $p = 2$, suggests that overestimating the lag (FDLM-FAR(3)) is preferable to underestimating the lag (FDLM-FAR(1), GP-FAR(1)) for ψ_1 estimation. FDLM-FAR(1) is competitive with GP-FAR(1), even when the parametric Gaussian process model assumes the correct (smooth) innovation distribution, which suggests that the FDLM implementation of Section 4.3 provides an adequate approximation. Under the dense design (see Appendix C), the improvements of the proposed methods over existing functional

data methods are less substantial, and for $T = 350$ the functional data methods all nearly achieve the oracle performance. The proposed methods, however, again provide superior recovery of ψ_1. In general, we find that the functional data methods, in particular the proposed approaches, outperform the multivariate methods, especially in the dense design. We conclude that the proposed methods provide highly competitive forecasts and superior FAR kernel recovery in a wide variety of important settings.

4.7 Forecasting Nominal and Real Yield Curves

We apply the proposed methods to model and forecast nominal and real yield curves. Yield curves are important in a variety of economic and financial applications, such as evaluating economic and monetary conditions, pricing fixed-income securities, generating forward curves, computing inflation premiums, and monitoring business cycles (Bolder et al., 2004). In practice, the U.S. real yield curve is estimated using Treasury Inflation-Protected Securities (TIPS), for which payments are adjusted according to the Consumer Price Index for All Urban Consumers (CPI-U) to provide investors with protection against inflation. U.S. nominal and TIPS yield curve data are published daily by the Federal Reserve, which uses actively-traded securities to fit a quasi-cubic spline for each curve. Estimates of the real and nominal yield curves are provided for maturities $\mathcal{T}_t^R = \{60, 84, 120, 240, 360\}$ and $\mathcal{T}_t^N = \{1, 3, 6, 12, 24, 36\} \cup \mathcal{T}_t^R$ months, respectively. Notably, the real yield is observed sparsely, and only at longer maturities. The small number of available maturities for real yields presents a challenge for existing functional time series models, and provides an interesting comparison with the nominal yield, for which there are more observed

maturities.

To assess the performance of the proposed model, we conducted an extensive forecasting study using daily nominal and real yield curve data. Beginning in 2003, we construct nine consecutive yet non-overlapping 18-month subperiods for estimation ($T \approx 375$); the corresponding starting dates are given in Table 4.1. For the month following each estimation period, we compute both one- and five-step (i.e., one business week) forecasts (≈ 20 and ≈ 15 time points, respectively) for *both* the nominal and real yields. In all cases, the nominal and real yields are modeled separately in order to provide additional comparisons.

We compute forecasts for the proposed methods by simulating from the forecasting distribution in the DLM (4.6). For computational convenience, we update only the DLM state parameters $\{\boldsymbol{\mu}_t\}$ during the forecast periods, and fix the remaining parameters based on the estimation periods. We also rescale the observation points $\mathcal{T}_t^R$ and $\mathcal{T}_t^N$ such that $\mathcal{T}_t^R, \mathcal{T}_t^N \subset \mathcal{T} = [0, 1]$. We compute forecasts using the competing methods described in Section 4.6, which use all available data for each forecast. For further comparisons, we include two popular parametric yield curve models based on the Nelson-Siegel parametrization (Nelson and Siegel, 1987): Diebold and Li (2006, DL), which extends the Nelson-Siegel model to the dynamic setting via a two-step estimation procedure, and Diebold et al. (2006, DRA) which is similar to DL, but instead estimates parameters jointly using maximum likelihood within a state space model; see Appendix C for implementation details.

The one- and five-step root mean squared forecasting errors (RMSFEs) for the nominal yields and real yields are in Tables 4.1 and 4.2, respectively. We omit unstable DRA forecasts, as well as multi-step forecasts for FAR Classic, which

are unavailable. For both data sets, the proposed methods—denoted FAR(1) and FAR(p), using the lag selection procedure with $p_{max} = 3$—are consistently among the best forecasters for all time periods, and outperform the existing functional data forecasts by a wide margin. For the nominal yields, the FAR(1) provides the best one-step forecasts aggregated across all time periods. For the real yields, the proposed methods are again among the most competitive, particularly in the periods since the financial crisis. Echoing the results in Diebold and Li (2006), the RW forecast is a difficult benchmark to clear, and the existing functional data models typically fail to do so. By comparison, the proposed FAR forecasts are highly competitive across all time periods and for both the nominal and (sparsely-observed) real yields.

An important feature of the proposed FAR model is the ability to compute exact (up to MCMC error) credible bands for parameters of interest, including forecasts. Such uncertainty quantification is unavailable for the RW forecast, which is our primary competitor in this application. For illustration, we compute pointwise and simultaneous credible bands for one-step forecasts during August 2016 in Figure 4.4. For both nominal and real yields, the credible bands are tighter for shorter maturities and widen in regions of unobserved points, which is appropriate behavior for a nonparametric method.

4.8 Concluding Remarks

The proposed hierarchical FAR(p) model provides a useful framework for estimation, inference, and forecasting functional time series data. Our model is especially suited for sparsely or irregularly sampled curves and for curves

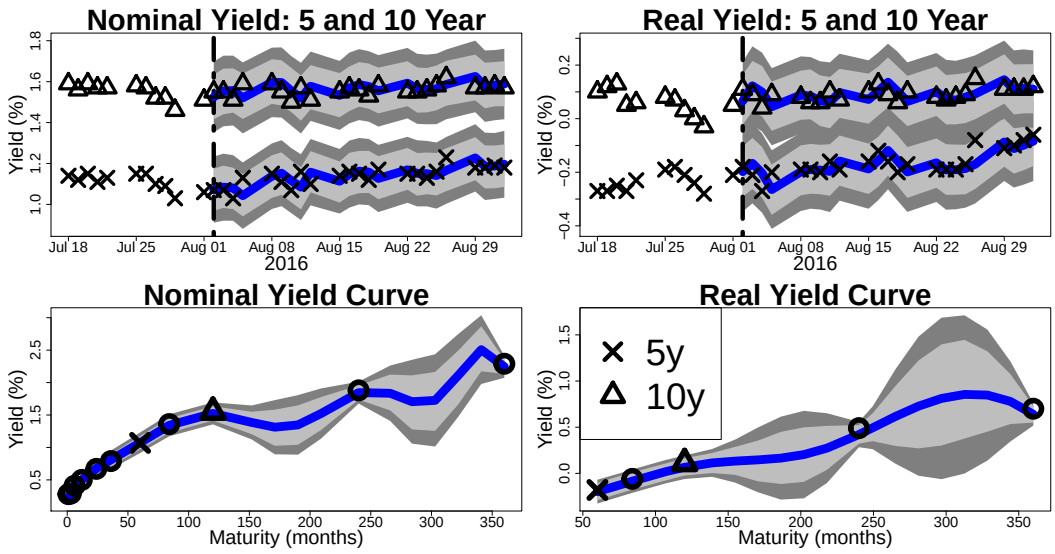


Figure 4.4: One-step nominal (**left**) and real (**right**) yield curve forecasts during 2016. **Top:** Time series of five ($\times$) and ten ($\triangle$) year observed maturities with one-step forecasts. **Bottom:** Observed (points) and forecast (line) curves on 8/2/16, corresponding to the dotted vertical line in the top panels. Posterior means (blue) and 95% pointwise and simultaneous prediction bands (light gray and dark gray, respectively) estimated using 10,000 MCMC simulations after a burn-in of 5,000.

sampled with non-negligible measurement error, and produces best linear predictors in a general FAR(p) setting, thereby dominating many competing functional time series models. The FDLM provides a more flexible, computationally efficient, and stable approach for modeling (innovation) covariance functions. Our model averaging procedure provides an effective solution to the problem of specifying p, and produces highly competitive forecasts. The simulation analysis and yield curve application suggest that the proposed FAR(p) model may improve forecasting and estimation in a wide range of settings, and the efficient MCMC sampling algorithm allows us to perform exact (up to MCMC error and prior misspecification) inference for important parameters.

While we assumed independent factors (and therefore independent innova-

tions) in Section 4.3, we can relax this assumption and allow $\boldsymbol{\Sigma}_e$ to be a stochastic process evolving over time. In this more general framework, the FDLM (4.7) can accommodate stochastic volatility or heavier-tailed distributions for the factors, yet retains the computational simplifications of (4.8) and Theorem 4.2. Letting $\boldsymbol{\Sigma}_t = \text{diag}\left(\{\sigma^2_{j,t}\}_{j=1}^{J_\epsilon}\right)$, the (time-dependent) innovation covariance function is $K_{\epsilon_t}(\tau, u) \equiv \text{Cov}\left(\epsilon_t(\tau), \epsilon_t(u)\right) = \sum_{j=1}^{J_\epsilon} \sigma^2_{j,t}\phi_j(\tau)\phi_j(u) + \sigma^2_\eta \mathbf{1}\{\tau = u\}$. By modeling each $\{\sigma^2_{j,t}\}_{t=1}^T$ for $j = 1, \ldots, J_\epsilon$ with an independent stochastic volatility model (e.g., Kim et al., 1998), the time-dependence of $\{\sigma^2_{j,t}\}$ will propagate to the innovation covariance functions, K_{ϵ_t}. Similar modifications can accommodate scale-mixtures of Gaussian distributions for the factors (Fernandez and Steel, 2000) to induce more general distributions for the innovation process, $\{\epsilon_t\}$. These generalizations are particularly important for financial applications, for which stochastic volatility models and heavy-tailed distributions are commonly appropriate.

Future work will investigate more adaptive FAR(p) models for longer, possibly nonstationary functional time series through stochastic volatility, time-varying ψ_ℓ, and regime shifts. Important extensions also include modeling multiple functional responses $Y_t(\tau) \in \mathbb{R}^d$ for $d > 1$, which requires a model for both the auto- and cross-correlations, and incorporating exogenous predictors. In both cases, the DLM framework of (4.6) offers a promising platform for pursuing these extensions.

Nominal Yields: h-Step Root Mean Squared Forecast Errors (RMSFEs)

	h	RW	Mean	VAR-Y	DL	DRA	FAR Classic	VAR-FPC	FAR(1)	FAR(p)
2/03	1	0.0488	0.4554	*0.0487*	0.1218	0.1440	0.1641	0.1631	0.0498	0.0516
	5	0.0966	0.4369	0.0904	0.1409	0.8221	-	0.1941	*0.0879*	0.1002
8/04	1	0.0253	1.1079	*0.0252*	0.0877	-	0.1113	0.1127	0.0281	0.0281
	5	0.0525	1.1279	*0.0383*	0.0953	-	-	0.1435	0.0412	0.0505
2/06	1	0.1710	0.5408	0.1809	0.2206	-	0.3349	0.3334	0.1682	*0.1673*
	5	*0.4534*	0.5971	0.5885	0.4927	-	-	0.5928	0.4680	0.4627
8/07	1	0.0833	1.3125	0.0860	0.1817	0.1854	0.1168	0.1173	0.0806	*0.0793*
	5	0.1345	1.3146	0.1402	0.2099	0.2998	-	0.1292	0.1537	*0.1233*
2/09	1	*0.0487*	0.5268	0.0517	0.1376	0.0917	0.1406	0.1398	0.0488	0.0760
	5	*0.0894*	0.5560	0.1227	0.1872	0.1451	-	0.1990	0.1323	0.2608
8/10	1	0.0344	0.5063	0.0333	0.1920	0.0878	0.0551	0.0554	*0.0291*	0.0292
	5	0.0583	0.4999	0.0603	0.1950	0.1356	-	0.0724	*0.0452*	0.0495
2/12	1	0.0383	0.5329	0.0384	0.0953	0.1915	0.0464	0.0463	0.0312	*0.0311*
	5	0.0951	0.5522	0.0915	0.1240	0.2476	-	0.0989	0.0760	*0.0734*
8/13	1	0.0463	0.4169	*0.0443*	0.0621	0.0692	0.0634	0.0644	0.0547	0.0676
	5	0.1210	0.3842	0.1104	0.1423	0.1448	-	*0.1100*	0.1208	*0.1100*
2/15	1	0.0329	0.3085	0.0320	0.1125	0.1001	0.0594	0.0606	*0.0305*	0.0321
	5	0.0420	0.3080	0.0403	0.1149	0.1202	-	0.0697	*0.0393*	0.0441

Table 4.1: h-step RMSFEs for nominal yields, grouped (left to right) by multivariate methods, parametric yield curve models, existing functional data methods, and proposed hierarchical FAR methods. The minimum RMSFE in each row is italicized.

Real Yields: h-Step Root Mean Squared Forecast Errors (RMSFEs)

	h	RW	Mean	VAR-Y	DL	DRA	FAR Classic	VAR-FPC	FAR(1)	FAR(p)
2/03	1	*0.0490*	0.1629	0.0504	0.0499	0.0492	0.1366	0.1329	0.0509	0.0572
	5	0.1001	0.1585	0.1040	0.1017	0.1128	-	0.1525	*0.0967*	0.1110
8/04	1	0.0331	0.3827	0.0337	0.0353	0.0528	0.0431	0.0440	0.0331	*0.0326*
	5	0.0724	0.3924	0.0707	0.0792	0.1690	-	0.0721	0.0679	*0.0651*
2/06	1	0.0429	0.1089	0.0428	0.0448	0.0453	0.0529	0.0533	*0.0424*	*0.0424*
	5	0.0934	0.1082	0.0858	0.0957	0.1362	-	0.0920	0.0852	*0.0835*
8/07	1	*0.0802*	0.2150	0.0896	0.0944	0.1979	0.1212	0.1202	0.0898	0.0880
	5	*0.1866*	0.2309	0.2268	0.2504	1.1843	-	0.1916	0.2051	0.1980
2/09	1	*0.0519*	0.5162	0.0544	0.0643	0.1229	0.0736	0.0749	0.0526	0.0541
	5	*0.0798*	0.5262	0.1100	0.1092	0.3606	-	0.1092	0.0992	0.1046
8/10	1	0.0490	0.7836	0.0492	0.0591	0.0663	0.0800	0.0762	0.0488	*0.0486*
	5	0.0735	0.7845	0.0787	0.0794	0.1815	-	0.0959	*0.0727*	0.0744
2/12	1	*0.0602*	0.8838	0.0612	0.0675	0.1492	0.0906	0.0853	0.0610	0.0608
	5	0.1845	0.9250	0.1958	0.1897	1.7442	-	0.2034	*0.1840*	0.1846
8/13	1	0.0526	0.3242	0.0506	0.0736	-	0.0613	0.0610	0.0500	*0.0492*
	5	0.1551	0.2981	0.1278	0.1380	-	-	0.1246	0.1407	*0.1239*
2/15	1	0.0328	0.3088	0.0327	0.0439	0.1529	0.0776	0.0779	*0.0325*	0.0336
	5	0.0489	0.3104	0.0521	0.0562	-	-	0.0816	*0.0466*	0.0543

Table 4.2: h-step RMSFEs for real yields, grouped (left to right) by multivariate methods, parametric yield curve models, existing functional data methods, and proposed hierarchical FAR methods. The minimum RMSFE in each row is italicized.

CHAPTER 5

CONCLUSIONS

The proposed methods provide effective Bayesian approaches for modeling functional and time series data. While broadly applicable, the proposed methodology directly addresses the following challenging cases for which existing methods are inadequate:

1. Functional data with additional complex dependence, such as time dependence, contemporaneous dependence, stochastic volatility, covariates, and change points (Chapter 2);
2. Functional data, time series data, or regression functions with local features, such as jumps or rapidly-changing smoothness (Chapter 3); and
3. Forecasting and inference of functional time series data with sparsely or irregularly sampled curves and for curves sampled with non-negligible measurement error (Chapter 4).

Using the MFDLM of Chapter 2, we may adapt general scalar and multivariate methods to the functional data setting. In particular, by separating out the functional component through appropriate conditioning and include the necessary identifiability constraints, the remaining dependence structures, such as covariates, repeated measurements, and spatial dependence, may be modeled via the factors. The hierarchical Bayesian approach allows us to incorporate interesting and useful submodels seamlessly, with minimal adjustments to the proposed Gibbs sampling algorithm.

An interesting extension of the MFDLM of Chapter 2 would be to incorporate the dynamic shrinkage processes of Chapter 3 to provide adaptive

shrinkage and regularization of the dynamic factors. Dynamic shrinkage processes inherit the desirable shrinkage behavior of global-local priors, such as the horseshoe prior, but with greater time-localization. By construction, the MFDLM—and more broadly, dynamic linear models—contains many parameters, and therefore may benefit from structured regularization. By synthesizing the MFDLM of Chapter 2 and the dynamic shrinkage processes of Chapter 3, we may model additional dependence among functional data, such as covariates, repeated measurements, and spatial dependence, while simultaneously introducing temporally adaptive shrinkage behavior to guard against overparametrization.

Important extensions of the dynamic shrinkage processes of Chapter 3 include alternative dependence models in (3.2) or multivariate shrinkage in (3.10). For example, dynamic shrinkage processes may offer effective shrinkage behavior for spatial or spatio-temporal models, in which case (3.2) may be modified to incorporate spatial dependence. Similarly, for replicate time series or functional data, multi-level extensions of (3.2) may provide both hierarchical and locally adaptive shrinkage behavior.

The proposed hierarchical FAR(p) model in Chapter 4 may be extended to incorporate exogenous predictors, stochastic volatility, time-varying autoregressive kernels ψ_ℓ, and regime shifts. These important generalizations may provide broader applicability of the hierarchical FAR(p) model for longer, possibly nonstationary functional time series data. As with the MFDLM, the dynamic linear model framework offers a promising platform for pursuing these extensions, and may be combined with dynamic shrinkage processes for additional locally adaptive regularization.

APPENDIX A

A BAYESIAN MULTIVARIATE FUNCTIONAL DYNAMIC LINEAR MODEL

To sample from the joint posterior distribution, we use a Gibbs sampler. Because the Gibbs sampler allows blocks of parameters to be conditioned on all other blocks of parameters, it is a convenient approach for our model. First, hierarchical dynamic linear model (DLM) algorithms typically require that $\boldsymbol{\beta}_t$ and $\boldsymbol{\theta}_t$ be the only unknown components, which we can accommodate by conditioning appropriately. Second, our sequential orthonormality approach for $f_k^{(c)}$ fits nicely within a Gibbs sampler, and we can adapt the algorithms described in Wand and Ormerod (2008). And third, the hierarchical structure of our model imposes natural conditional independence assumptions, which allows us to easily partition the parameters into appropriate blocks.

A.1 Initialization

To initialize the factors $\boldsymbol{\beta}_k^{(c)} = \left(\beta_{k,1}^{(c)}, \ldots, \beta_{k,T}^{(c)}\right)'$ and the factor loading curves (FLCs) $f_k^{(c)}$ for $k = 1, \ldots, K$ and $c = 1, \ldots, C$, we compute the singular value decomposition (SVD) of the data matrix $\mathbf{Y}^{(c)} = \mathbf{U}^{(c)}\boldsymbol{\Sigma}^{(c)}\mathbf{V}^{(c)\prime}$ for $c = 1, \ldots, C$. Note that to obtain a data *matrix* $\mathbf{Y}^{(c)}$, with rows corresponding to times t and columns to observations points τ, we need to estimate $Y_t^{(c)}(\tau)$ for any unobserved τ at each time t, which may be computed quickly using splines. However, these estimated data values are *only* used for the initialization step. Then, letting $\mathbf{U}_{1:K}^{(c)}$ be the first K columns of $\mathbf{U}^{(c)}$, $\boldsymbol{\Sigma}_{1:K}^{(c)}$ be the upper left $K \times K$ submatrix of $\boldsymbol{\Sigma}^{(c)}$, and $\mathbf{V}_{1:K}^{(c)}$ be the first K columns of $\mathbf{V}^{(c)}$, we initialize the factors $\left(\boldsymbol{\beta}_1^{(c)}, \ldots, \boldsymbol{\beta}_K^{(c)}\right) = \mathbf{U}_{1:K}^{(c)}\boldsymbol{\Sigma}_{1:K}^{(c)}$ and the FLCs $\left(\boldsymbol{f}_1^{(c)}, \ldots, \boldsymbol{f}_K^{(c)}\right) = \mathbf{V}_{1:K}^{(c)}$, where $\boldsymbol{f}_k^{(c)}$

is the vector of FLC k evaluated at all observation points $\cup_t \mathcal{T}_t^{(c)}$ for outcome c. The $\boldsymbol{f}_k^{(c)}$ are orthonormal in the sense that $\boldsymbol{f}_k^{(c)\prime}\boldsymbol{f}_j^{(c)} = \mathbf{1}(k = j)$, but they are not smooth. This approach is similar to the initializations in Matteson et al. (2011) and Hays et al. (2012).

Given the factors $\boldsymbol{\beta}_k^{(c)}$ and the FLCs $\boldsymbol{f}_k^{(c)}$, we can estimate each $\sigma^2_{(c)}$ (or more generally, $\mathbf{E}_t$) as a conditional maximum likelihood estimator (MLE), using the likelihood from the observation level of model (2.1). Similarly, we can estimate each $\lambda_k^{(c)}$ conditional on $\boldsymbol{f}_k^{(c)}$ by maximizing the partially informative normal likelihood. Then, given $\lambda_k^{(c)}$, $\sigma^2_{(c)}$, $\boldsymbol{\beta}_k^{(c)}$, and $\boldsymbol{f}_k^{(c)}$, we can estimate each $\boldsymbol{d}_k^{(c)}$ by normalizing the full conditional posterior expectation given in the main paper; i.e., solving the relevant quadratic program and then normalizing the solution. Initializations for the remaining levels proceed similarly as conditional MLEs, but depend on the form chosen for $\mathbf{X}_t$, $\mathbf{V}_t$, $\mathbf{G}_t$, and $\mathbf{W}_t$. In our applications, this conditional MLE approach produces reasonable starting values for all variables.

A.1.1 Common Factor Loading Curves

If we wish to implement the common FLCs model $f_k^{(c)} = f_k$ for all k, c, then we instead compute the SVD of the stacked data matrices $\left(\mathbf{Y}^{(1)\prime}, \ldots, \mathbf{Y}^{(C)\prime}\right)' = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}'$, where now the data matrices $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(C)}$ are imputed using splines for all observation points for all outcomes, $\cup_{t,c}\mathcal{T}_t^{(c)}$, and therefore have the same number of columns. Alternatively, we may improve computational efficiency by choosing a small yet representative subset of observation points $\mathcal{T}^* \subset \cup_{t,c}\mathcal{T}_t^{(c)}$ and then estimating each data matrix $\mathbf{Y}^{(c)}$ for all $\tau \in \mathcal{T}^*$. Let $\mathbf{U}^{(c)}_{1:K}$ be the first K columns of $\mathbf{U}^{(c)}$, where the $\mathbf{U}^{(c)}, c = 1, \ldots, C$, correspond to the outcome-

specific blocks of $\mathbf{U} = \left(\mathbf{U}^{(1)\prime}, \ldots, \mathbf{U}^{(C)\prime}\right)'$. Then, similar to before, we set $\left(\boldsymbol{\beta}_1^{(c)}, \ldots, \boldsymbol{\beta}_K^{(c)}\right) = \mathbf{U}_{1:K}^{(c)}\boldsymbol{\Sigma}_{1:K}$ for $c = 1, \ldots, C$, and $(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_K) = \mathbf{V}_{1:K}$, where $\boldsymbol{\Sigma}_{1:K}$ is the upper left $K \times K$ submatrix of $\boldsymbol{\Sigma}$ and $\mathbf{V}_{1:K}$ is the first K columns of $\mathbf{V}$. Again, the $\boldsymbol{f}_k$ are unsmoothed with $\boldsymbol{f}_k'\boldsymbol{f}_j = \mathbf{1}(k = j)$, but now the initialized FLCs are common for $c = 1, \ldots, C$. Initialization of the remaining parameters proceeds as before, but now with $\lambda_k^{(c)} = \lambda_k$ and $\boldsymbol{d}_k^{(c)} = \boldsymbol{d}_k$, which can be obtained by maximizing the relevant conditional likelihoods under the common FLCs model.

A.2 Sampling

A.2.1 General Algorithm

For greater generality, we present our sampling algorithm for non-common FLCs; i.e., we retain dependence on c for $\boldsymbol{d}_k^{(c)}$ and $\lambda_k^{(c)}$. When applicable, we discuss the necessary modifications for the common FLCs model.

The algorithm proceeds in four main blocks:

1. Sample the basis coefficients $\boldsymbol{d}_k^{(c)}$ and the smoothing parameters $\lambda_k^{(c)}$ for the FLCs. For $\lambda_k^{(c)}$, we use a Gamma(γ_1, γ_2) prior distribution, which is conjugate to the partially informative normal likelihood and implies that the full conditional posterior distribution is Gamma$(\gamma_1 + \text{rank}(\boldsymbol{\Omega}_\phi)/2, \gamma_2 + \boldsymbol{d}_k^{(c)\prime}\boldsymbol{\Omega}_\phi\boldsymbol{d}_k^{(c)}/2)$. For the common FLCs model, we simply replace $\boldsymbol{d}_k^{(c)}$ with $\boldsymbol{d}_k$ to obtain the full conditional posterior for λ_k. We use the hyperparameters $\gamma_1 = \gamma_2 = 0.001$, although the effect of the hyperparameters is negligi-

ble as long as γ_1 and γ_2 are small relative to $\text{rank}(\boldsymbol{\Omega}_\phi)/2$ and $\boldsymbol{d}_k^{(c)\prime}\boldsymbol{\Omega}_\phi\boldsymbol{d}_k^{(c)}/2$, respectively. After sampling the $\lambda_k^{(c)}$, we sample and then normalize the $\boldsymbol{d}_k^{(c)}$ with a modified version of the efficient Cholesky decomposition approach of Wand and Ormerod (2008):

(a) Compute the (lower triangular) Cholesky decomposition $\mathbf{B}_k^{-1} = \bar{\mathbf{B}}_L\bar{\mathbf{B}}_L'$;

(b) If $k = 1$, set $\mathbf{L}'_{1:(k-1)}\boldsymbol{\Lambda}_{1:(k-1)} = \mathbf{0}$;

If $k > 1$, use forward substitutions to obtain $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ from the equations $\bar{\mathbf{B}}_L\bar{\mathbf{x}} = \mathbf{L}'_{1:(k-1)}$ and $\bar{\mathbf{B}}_L\bar{\mathbf{y}} = \mathbf{b}_k$, and let $\boldsymbol{\Lambda}_{1:(k-1)}$ be the solution to the regression of $\bar{\mathbf{y}}$ on $\bar{\mathbf{x}}$;

(c) Use forward substitution to obtain $\bar{\mathbf{b}}$ as the solution to $\bar{\mathbf{B}}_L\bar{\mathbf{b}} = \mathbf{b}_k$, then use backward substitution to obtain $\boldsymbol{d}_k^*$ as the solution to $\bar{\mathbf{B}}_L'\boldsymbol{d}_k^* = \bar{\mathbf{b}} + \bar{\mathbf{z}}$, where $\bar{\mathbf{z}} \sim N(\mathbf{0}, \mathbf{I}_{(M+4)\times(M+4)})$;

(d) Retain the vector $\boldsymbol{d}_k^{(c)} = \boldsymbol{d}_k^*/\sqrt{\boldsymbol{d}_k^{*\prime}\mathbf{J}_\phi\boldsymbol{d}_k^*}$ and set $\boldsymbol{\beta}_k^{(c)} = \sqrt{\boldsymbol{d}_k^{*\prime}\mathbf{J}_\phi\boldsymbol{d}_k^*}\boldsymbol{\beta}_k^{(c)}$.

The definitions of $\mathbf{B}_k$ and $\mathbf{b}_k$ depend on whether or not we use the common FLCs model with $f_k^{(c)} = f_k$. Compared with unconstrained Bayesian splines, the extra orthogonality step (b) uses the Cholesky decomposition—which we must compute regardless—and adds only the computational cost of a simple linear regression for each $k > 1$, which is perhaps expected in light of Theorem 2.1. The scaling of $\boldsymbol{d}_k^{(c)}$ and $\boldsymbol{\beta}_k^{(c)}$ in (d) enforces the unit-norm constraint on $f_k^{(c)}$ yet ensures that $f_k^{(c)}(\tau)\boldsymbol{\beta}_k^{(c)}$—which appears in the posterior distribution of $\boldsymbol{d}_j^{(c)}$ for all $j \neq k$—is unaffected by the normalization.

2. Sample the factors $\boldsymbol{\beta}_t$ (and $\boldsymbol{\theta}_t$, if present) conditional on all other parameters in (2.1) using either the DLM implementation of *forward filtering back-*

ward sampling (e.g., Petris et al. (2009)) or the state space sampler of Durbin and Koopman (2002); Koopman and Durbin (2003, 2000), the latter of which is optimized when $\mathbf{E}_t$ is diagonal. For general hierarchical models, we may modify the hierarchical DLM algorithms of Gamerman and Migon (1993).

For the prior distributions, we only need to specify the distribution of $\boldsymbol{\beta}_0$ (and $\boldsymbol{\theta}_0$); the remaining distributions are computed recursively using $\mathbf{F}$, $\mathbf{X}_t$, $\mathbf{G}_t$ and the error variances. For simplicity, we let $\beta_{k,0}^{(c)} \overset{iid}{\sim} N(0, 10^6)$, which is a common choice for DLMs. Alternatively, we could use past data not included in our analysis to estimate these initial values. However, the resulting estimates for $t > 1$ in our applications are not noticeably different.

3. Sample the state evolution matrix $\mathbf{G}_t$ (if unknown). $\mathbf{G}_t$ may have a special form (see Section A.2.2) or provide a more common time series model such as a VAR. In the latter case, we may choose some structure for $\mathbf{G}_t = \mathbf{G}$, e.g. diagonality to allow dependence between $\beta_{k,t}^{(c)}$ and $\beta_{k,t-1}^{(c)}$, or K blocks of dimension $C \times C$ to allow dependence between $\beta_{k,t}^{(c)}$ and $\beta_{k,t-1}^{(c')}$ for $c, c' = 1, \ldots, C$. A simple choice of prior for the nonzero entries of $\mathbf{G}$ is iid $N(0, 10^6)$, which is conjugate to the likelihood induced by (2.1). Under this prior, it is straightforward to derive the posterior distribution of $\text{vec}_0(\mathbf{G})$, where vec_0 stacks the nonzero entries of the matrix (by column) into a vector.

4. Sample each of the remaining error variance parameters individually: $\mathbf{E}_t$, $\mathbf{V}_t$, and $\mathbf{W}_t$. These distributions depend on our assumptions for the model structure, but we typically prefer conjugate priors when available. For example, in the random walk factor model of (8), we have

$\boldsymbol{\beta}_{k,i,s,t} = \boldsymbol{\beta}_{k,i,s,t-1} + \boldsymbol{\omega}_{k,i,s,t}$ with $\boldsymbol{\omega}_{k,i,s,t} \overset{indep}{\sim} N(\mathbf{0}, \mathbf{W}_k)$. Using the Wishart prior $\mathbf{W}_k^{-1} \sim$ Wishart$((\rho R)^{-1}, \rho)$, the full conditional posterior distribution for the precision is $\mathbf{W}_k^{-1} \sim$ Wishart$((\rho R + \sum_{i,s,t} \mathbf{w}_{k,i,s,t}\mathbf{w}'_{k,i,s,t})^{-1}, \rho + T)$, where $\mathbf{w}_{k,i,s,t} = \boldsymbol{\beta}_{k,i,s,t} - \boldsymbol{\beta}_{k,i,s,t-1}$ is conditional on the factors and $T = (15)(40)(8) = 4800$ counts the indices (i, s, t). We let $R^{-1} = \mathbf{I}_{C\times C}$, which is the expected prior precision, and $\rho = C \geq \text{rank}(R^{-1})$.

For the stochastic volatility model of Section 2.4.1, we use the distributions given in Kim et al. (1998). In particular, letting $\sigma^2_{k,(c),t} = \exp(h^{(c)}_{k,t})$, Kim et al. (1998) propose the model $h^{(c)}_{k,t} = \xi^{(c)}_{k,0} + \xi^{(c)}_{k,1}(h^{(c)}_{k,t-1} - \xi^{(c)}_{k,0}) + \zeta^{(c)}_{k,t}$, where $\zeta^{(c)}_{k,t} \overset{indep}{\sim} N(0, \sigma^2_{H,k,(c)})$ for $t = 2, \ldots, T$ and $h^{(c)}_{k,1} \sim N(\xi^{(c)}_{k,0}, \sigma^2_{H,k,(c)}/(1 - (\xi^{(c)}_{k,1})^2))$ with $|\xi^{(c)}_{k,1}| < 1$ for stationarity. Kim et al. (1998) also suggest priors for $\xi^{(c)}_{k,0}, \xi^{(c)}_{k,1}$, and $\sigma^2_{H,k,(c)}$ and provide an efficient MCMC sampling algorithm. For additional motivation for the stochastic volatility approach over GARCH models, see Daníelsson (1998).

Recall that we construct a posterior distribution of $\boldsymbol{d}^{(c)}_k$ without the unit norm constraint, and then normalize the samples from this distribution. As a result, the conditions of Theorem 2.1 are satisfied and the (unnormalized) full conditional posterior distribution of $\boldsymbol{d}^{(c)}_k$ is Gaussian, both of which are convenient results. The normalization step 1.(d) is interpretable, corresponding to the projection of a Gaussian distribution onto the unit sphere. Note that rescaling the factors $\boldsymbol{\beta}^{(c)}_k$ in 1.(d) does not affect the remainder of the sampling algorithm (steps 2. - 4.). The rescaled $\boldsymbol{\beta}^{(c)}_k$ are from the previous MCMC iteration, which does not affect the full conditional distributions of step 2. in the current MCMC iteration. The subsequent steps 3., 4., and 1. are then conditional on the newly sampled factors $\boldsymbol{\beta}^{(c)}_k$ from step 2., which have not been rescaled.

A.2.2 Sampling the Common Trend Hidden Markov Model

Recall the common trend hidden Markov model for the factors, $k = 1, \ldots, K$:

$$\begin{cases} \Delta^D \beta_{k,t}^{(1)} = \omega_{k,t}^{(1)}, & \omega_{k,t}^{(1)} = \sum_{i=1}^{r} \psi_{k,i}^{(1)} \omega_{k,t-i}^{(1)} + \sigma_{k,(1),t} z_{k,t}^{(1)} \\ \Delta^D \beta_{k,t}^{(c)} = s_{k,t}^{(c)} (\gamma_k^{(c)} \Delta^D \beta_{k,t}^{(1)}) + \omega_{k,t}^{(c)}, & \omega_{k,t}^{(c)} = \sum_{i=1}^{r} \psi_{k,i}^{(c)} \omega_{k,t-i}^{(c)} + \sigma_{k,(c),t} z_{k,t}^{(c)} \end{cases} \tag{A.1}$$

for $c = 2, \ldots, C$, where Δ is the differencing operator, D is the degree of differencing, $\gamma_k^{(c)} \in \mathbb{R}$ is the economy-specific slope term for each factor, $\left\{ s_{k,t}^{(c)} : t = 1 \ldots, T \right\}$ is a discrete Markov chain with states $\{0, 1\}$, $\sigma_{k,(c),t}^2$ are the time-dependent error variances, and $z_{k,t}^{(c)} \overset{iid}{\sim} N(0, 1)$. We specify iid $N(0, 10^6)$ priors for $\gamma_k^{(c)}$, which are conjugate to the likelihood in (A.1).

We can express (A.1) as the $\boldsymbol{\beta}_t = \boldsymbol{\theta}_t$-level in (2.1) with $\mathbf{X}_t = \mathbf{I}_{CK \times CK}$ and $\mathbf{V}_t = \mathbf{0}_{CK \times CK}$. Let $\boldsymbol{L}_{\boldsymbol{\beta}_t} = \mathbf{I}_{CK \times CK} - \mathbf{Q}_t$,

$$\mathbf{Q}_t = \begin{pmatrix} \mathbf{0}_{K \times K} & \mathbf{0}_{K \times K} & \cdots & \mathbf{0}_{K \times K} \\ \mathbf{S}_t^{(2)} \boldsymbol{\gamma}^{(2)} & \mathbf{0}_{K \times K} & \cdots & \mathbf{0}_{K \times K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_t^{(C)} \boldsymbol{\gamma}^{(C)} & \mathbf{0}_{K \times K} & \cdots & \mathbf{0}_{K \times K} \end{pmatrix},$$

where $\mathbf{S}_t^{(c)} = \text{diag}(\{s_{k,t}^{(c)}\}_{k=1}^K)$ and $\boldsymbol{\gamma}^{(c)} = \text{diag}(\{\gamma_k^{(c)}\}_{k=1}^K)$. Note that $\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1} = \mathbf{I}_{CK \times CK} + \mathbf{Q}_t$. To derive the state evolution matrix $\mathbf{G}_t$, we can modify the standard ARIMA(r, D,0) framework for DLMs to incorporate the $s_{k,t}^{(c)}$-dependent common trend. For example, when $D = r = 1$, we have $\Delta\beta_{k,t}^{(c)} - s_{k,t}^{(c)} \gamma_k^{(c)} \Delta\beta_{k,t}^{(1)} = \psi_k^{(c)} (\Delta\beta_{k,t-1}^{(c)} - s_{k,t-1}^{(c)} \gamma_k^{(c)} \Delta\beta_{k,t-1}^{(1)}) + \sigma_{k,(c),t} z_{k,t}^{(c)}$ which can be rewritten as $\beta_{k,t}^{(c)} - s_{k,t}^{(c)} \gamma_k^{(c)} \beta_{k,t}^{(1)} = (1 + \psi_k^{(c)}) \beta_{k,t-1}^{(c)} - (s_{k,t}^{(c)} + s_{k,t-1}^{(c)} \psi_k^{(c)}) \gamma_k^{(c)} \beta_{k,t-1}^{(1)} - \psi_k^{(c)} \beta_{k,t-2}^{(c)} + \psi_k^{(c)} s_{k,t-1}^{(c)} \gamma_k^{(c)} \beta_{k,t-2}^{(1)} + \sigma_{k,(c),t} z_{k,t}^{(c)}$. The left side of this equation is given by the elements of $\boldsymbol{L}_{\boldsymbol{\beta}_t} \boldsymbol{\beta}_t$, while the right side may clearly be expressed using a simple modification of the standard ARIMA DLM state evolution matrix $\mathbf{G}$. In vector

notation, we have

$$\begin{pmatrix} \boldsymbol{L}_{\boldsymbol{\beta}_t} & \mathbf{0}_{CK\times CK} \\ \mathbf{0}_{CK\times CK} & \mathbf{I}_{CK\times CK} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{\beta}_{t-1} \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{t,1} & \mathbf{G}_{t,2} \\ \mathbf{0}_{CK\times CK} & \mathbf{I}_{CK\times CK} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_{t-2} \end{pmatrix} + \begin{pmatrix} \tilde{\boldsymbol{\omega}}_t \\ \tilde{\boldsymbol{\omega}}_{t-1} \end{pmatrix} \tag{A.2}$$

where

$$\mathbf{G}_{t,1} = (\mathbf{I}_{CK\times CK} + \boldsymbol{\Psi}) + \begin{pmatrix} \mathbf{0}_{K\times K} & \mathbf{0}_{K\times(C-1)K} \\ -(\mathbf{S}_t^{(2)} + \mathbf{S}_{t-1}^{(2)}\boldsymbol{\Psi}^{(2)})\boldsymbol{\gamma}^{(2)} & \mathbf{0}_{K\times(C-1)K} \\ \vdots & \vdots \\ -(\mathbf{S}_t^{(C)} + \mathbf{S}_{t-1}^{(C)}\boldsymbol{\Psi}^{(C)})\boldsymbol{\gamma}^{(C)} & \mathbf{0}_{K\times(C-1)K} \end{pmatrix},$$

$$\mathbf{G}_{t,2} = -\boldsymbol{\Psi} + \begin{pmatrix} \mathbf{0}_{K\times K} & \mathbf{0}_{K\times(C-1)K} \\ \mathbf{S}_{t-1}^{(2)}\boldsymbol{\Psi}^{(2)}\boldsymbol{\gamma}^{(2)} & \mathbf{0}_{K\times(C-1)K} \\ \vdots & \vdots \\ \mathbf{S}_{t-1}^{(C)}\boldsymbol{\Psi}^{(C)}\boldsymbol{\gamma}^{(C)} & \mathbf{0}_{K\times(C-1)K} \end{pmatrix}, \text{ and}$$

$$\text{Var}\begin{pmatrix} \tilde{\boldsymbol{\omega}}_t \\ \tilde{\boldsymbol{\omega}}_{t-1} \end{pmatrix} = \begin{pmatrix} \mathbf{W}_t & \mathbf{0}_{CK\times CK} \\ \mathbf{0}_{CK\times CK} & \mathbf{0}_{CK\times CK} \end{pmatrix},$$

with $\boldsymbol{\Psi} = \text{diag}(\{\psi_k^{(c)}\}_{k,c})$, $\mathbf{W}_t = \text{diag}(\{\sigma^2_{k,(c),t}\}_{k,c})$, and $\tilde{\boldsymbol{\omega}}_t$ has elements $\tilde{\omega}_{k,t}^{(c)} = \sigma_{k,(c),t} z_{k,t}^{(c)}$, which are the residuals from the AR(r) process in (A.1).

Many of these matrix multiplications involve diagonal matrices, and therefore may be computed quickly. The error variance is not a proper variance matrix, but is commonly used for sampling DLMs with multiple lags or differencing. Note that to write (2.1) in this form, we must also append CK columns of zeros to $\mathbf{F}(\tau)$, since $\mathbf{Y}_t(\tau)$ depends on $\boldsymbol{\beta}_t$ but not on $\boldsymbol{\beta}_{t-1}$.

Inverting the block diagonal matrix $\tilde{\boldsymbol{L}}_{\boldsymbol{\beta}_t} = \text{bdiag}(\boldsymbol{L}_{\boldsymbol{\beta}_t}, \mathbf{I}_{CK\times CK})$, we obtain $\tilde{\boldsymbol{L}}_{\boldsymbol{\beta}_t}^{-1} = \text{bdiag}(\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1}, \mathbf{I}_{CK\times CK}) = \text{bdiag}(\mathbf{I}_{CK\times CK} + \mathbf{Q}_t, \mathbf{I}_{CK\times CK})$. Therefore, we

can rewrite (A.2) as

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{\beta}_{t-1} \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{t,1} + \mathbf{Q}_t\mathbf{G}_{t,1} & \mathbf{G}_{t,2} + \mathbf{Q}_t\mathbf{G}_{t,2} \\ \mathbf{0}_{CK\times CK} & \mathbf{I}_{CK\times CK} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{t-1} \\ \boldsymbol{\beta}_{t-2} \end{pmatrix} + \tilde{\boldsymbol{L}}_{\boldsymbol{\beta}_t}^{-1} \begin{pmatrix} \tilde{\boldsymbol{\omega}}_t \\ \tilde{\boldsymbol{\omega}}_{t-1} \end{pmatrix} \quad \text{(A.3)}$$

where the error variance has the same block form as previously, but with $\mathbf{W}_t$ replaced by $\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1}\mathbf{W}_t(\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1})' = (\mathbf{I}_{CK\times CK} + \mathbf{Q}_t)\mathbf{W}_t(\mathbf{I}_{CK\times CK} + \mathbf{Q}_t') = \mathbf{W}_t + \mathbf{Q}_t\mathbf{W}_t + (\mathbf{Q}_t\mathbf{W}_t)' + \mathbf{Q}_t\mathbf{W}_t\mathbf{Q}_t'$. Letting $\boldsymbol{\sigma}^2_{(c),t} = \text{diag}(\{\sigma^2_{k,(c),t}\}_{k=1}^K)$ so that $\mathbf{W}_t = \text{bdiag}(\boldsymbol{\sigma}^2_{(1),t}, \ldots, \boldsymbol{\sigma}^2_{(C),t})$, we may compute the relevant terms explicitly:

$$\mathbf{Q}_t\mathbf{W}_t = \begin{pmatrix} \mathbf{0}_{K\times K} & \mathbf{0}_{K\times K} & \cdots & \mathbf{0}_{K\times K} \\ \mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\sigma}^2_{(1),t} & \mathbf{0}_{K\times K} & \cdots & \mathbf{0}_{K\times K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t} & \mathbf{0}_{K\times K} & \cdots & \mathbf{0}_{K\times K} \end{pmatrix}$$

and

$$\mathbf{Q}_t\mathbf{W}_t\mathbf{Q}_t' = \begin{pmatrix} \mathbf{0}_{K\times K} & \mathbf{0}_{K\times K} & \cdots & \mathbf{0}_{K\times K} \\ \mathbf{0}_{K\times K} & \mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\sigma}^2_{(1),t}\mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)} & \cdots & \mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\sigma}^2_{(1),t}\mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{K\times K} & \mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t}\mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)} & \cdots & \mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t}\mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)} \end{pmatrix}$$

where again, the component terms are all diagonal, and therefore can be re-ordered for convenience. Combining terms and simplifying, the nonzero upper left block of the error variance matrix, $\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1}\mathbf{W}_t(\boldsymbol{L}_{\boldsymbol{\beta}_t}^{-1})'$, is

$$\begin{pmatrix} \boldsymbol{\sigma}^2_{(1),t} & \mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\sigma}^2_{(1),t} & \cdots & \mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t} \\ \mathbf{S}_t^{(2)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\sigma}^2_{(1),t} & \boldsymbol{\sigma}^2_{(2),t} + \mathbf{S}_t^{(2)}(\boldsymbol{\gamma}^{(2)})^2\boldsymbol{\sigma}^2_{(1),t} & \cdots & \mathbf{S}_t^{(2)}\mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t} & \mathbf{S}_t^{(2)}\mathbf{S}_t^{(C)}\boldsymbol{\gamma}^{(2)}\boldsymbol{\gamma}^{(C)}\boldsymbol{\sigma}^2_{(1),t} & \cdots & \boldsymbol{\sigma}^2_{(C),t} + \mathbf{S}_t^{(C)}(\boldsymbol{\gamma}^{(C)})^2\boldsymbol{\sigma}^2_{(1),t} \end{pmatrix}.$$

When $s_{k,t}^{(c)} = 1, c > 1$ the slope parameter $\gamma_k^{(c)}$ may increase or decrease the error variance of the residuals $\tilde{\omega}_{k,t}^{(c)}$ at time t, and determines the contemporaneous covariance between $\tilde{\omega}_{k,t}^{(c)}$ and $\tilde{\omega}_{k,t}^{(1)}$. Similarly, when $s_{k,t}^{(c)} = s_{k,t}^{(c')} = 1$, the

product $\gamma_k^{(c)}\gamma_k^{(c')}\sigma^2_{k,(1),t}$ determines the contemporaneous covariance between $\tilde{\omega}_{k,t}^{(c)}$ and $\tilde{\omega}_{k,t}^{(c')}$ at time t. Note that as long as there exist distinct times t, t' such that $s_{k,t}^{(c)} \neq s_{k,t'}^{(c)}$, the slopes and volatilities are identifiable for each k and c. Therefore, the common trend hidden Markov model of (A.1) provides a flexible, time-dependent contemporaneous covariance structure within a relatively simple regression framework.

A.3 Additional Figures

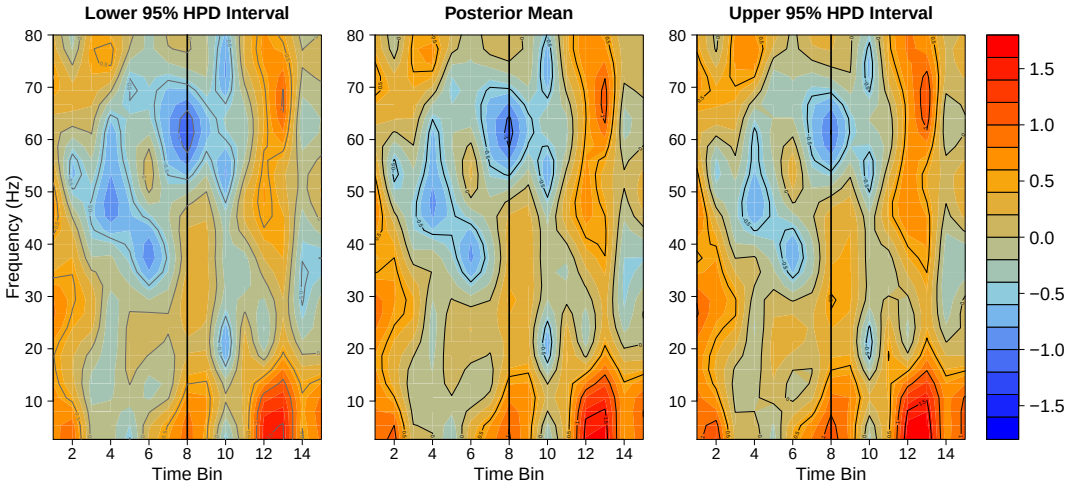


Figure A.1: Pointwise 95% HPD intervals and the posterior mean for $\bar{\mu}_t^{(1)}$, which is the average difference in the PFC log-spectra between the FC and FS trials. The black vertical lines indicate t^*.

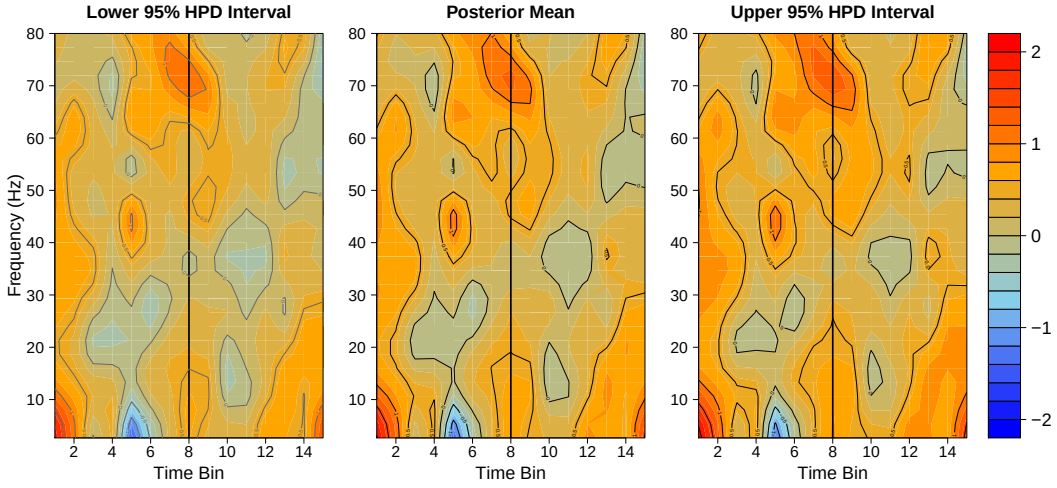


Figure A.2: Pointwise 95% HPD intervals and the posterior mean for $\bar{\mu}_t^{(2)}$, which is the average difference in the PFC log-spectra between the FC and FS trials. The black vertical lines indicate t^*.

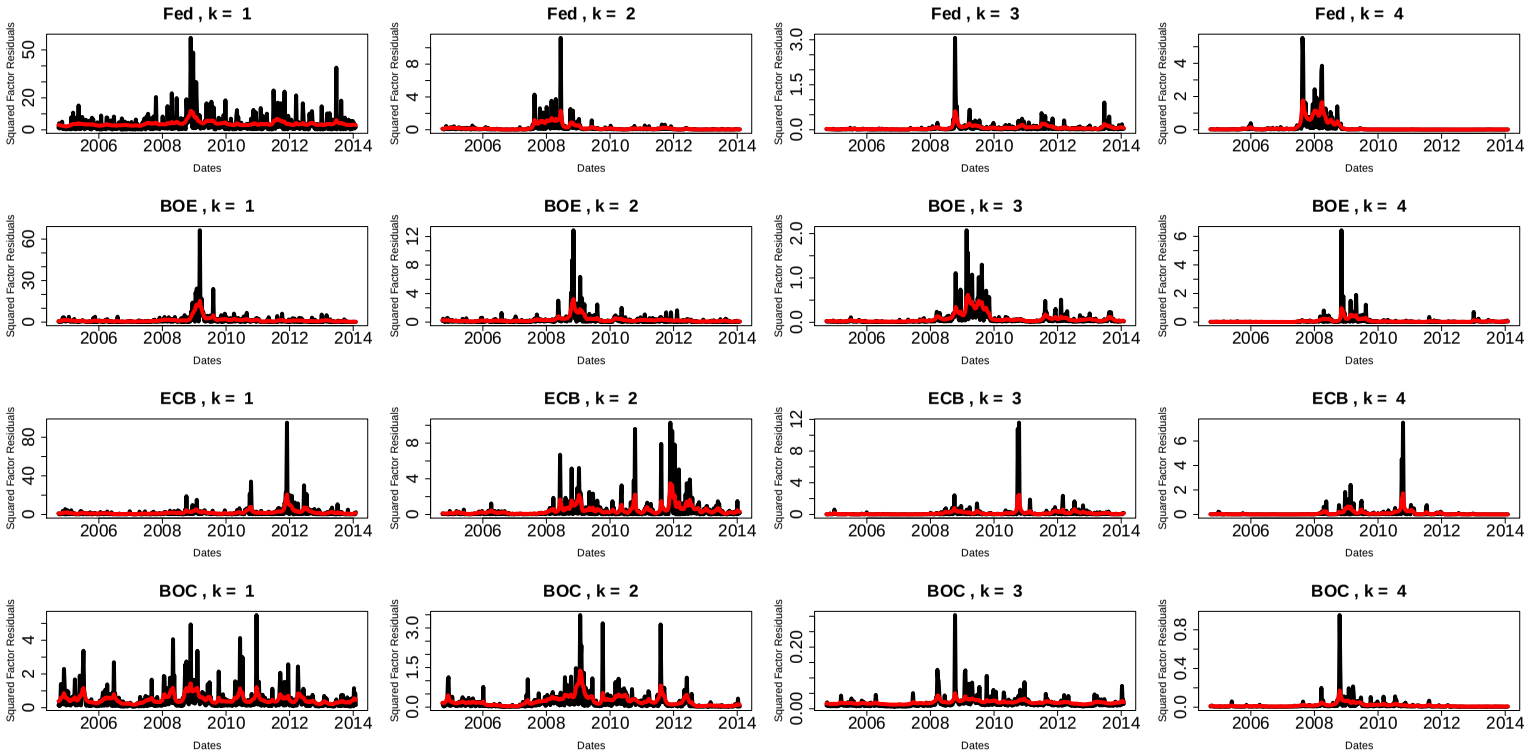


Figure A.3: The observed volatility clustering from the yield curve application. The black lines are the posterior means of the squared residuals from the AR(1) process on the $\omega_{k,t}^{(c)}$ in the common trend hidden Markov model of Section 2.4.1. The red lines are the posterior means of the corresponding volatility estimates $\sigma_{k,(c),t}^2$ discussed in Section 2.4.1.

APPENDIX B

DYNAMIC SHRINKAGE PROCESSES

Proof. (Proposition 3.1) Proposition 3.1 follows from Proposition 3.2 with $\mu_z = 0$. □

Proof. (Proposition 3.2) Let $\eta \sim Z(\alpha, \beta, \mu_z, 1)$ with density (3.3), i.e.,

$$[z] = \big[\sigma B(\alpha,\beta)\big]^{-1}\big\{\exp\big[(z-\mu_z)/\sigma_z\big]\big\}^{\alpha}\big\{1+\exp\big[(z-\mu_z)/\sigma_z\big]\big\}^{-(\alpha+\beta)}.$$

The density of $\lambda^2 = \exp(\eta)$ is

$$\begin{aligned}\big[\lambda^2\big] &\propto \big(\lambda^2\big)^{-1}\big\{\exp\big[\log(\lambda^2)-\mu_z\big]\big\}^{\alpha}\big\{1+\exp\big[\log(\lambda^2)-\mu_z\big]\big\}^{-(\alpha+\beta)} \\ &\propto \big(\lambda^2\big)^{\alpha-1}\big[1+\lambda^2/\exp(\mu_z)\big]^{-(\alpha+\beta)}\end{aligned}$$

and therefore the density of $\kappa = 1/(1+\lambda^2)$ is

$$\begin{aligned}[\kappa] &\propto \kappa^{-2}\big[\kappa^{-1}-1\big]^{\alpha-1}\big[1+(\kappa^{-1}-1)/\exp(\mu_z)\big]^{-(\alpha+\beta)} \\ &\propto \kappa^{-2-(\alpha-1)}(1-\kappa)^{\alpha-1}\left\{\kappa^{-1}\big[\kappa\exp(\mu_z)+(1-\kappa)\big]\right\}^{-(\alpha+\beta)} \\ &\propto (1-\kappa)^{\alpha-1}\kappa^{\beta-1}\big[\kappa\exp(\mu_z)+(1-\kappa)\big]^{-(\alpha+\beta)}\end{aligned}$$

i.e., $\kappa \sim \text{TPB}(\beta, \alpha, \exp(\mu_z))$. □

Proof. (Theorem 3.1) Under model (3.2), i.e.,

$$h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t, \quad \eta_t \overset{iid}{\sim} Z(\alpha, \beta, 0, 1),$$

we have $[h_{t+1}|h_t, \phi, \mu] \sim Z(\alpha, \beta, \mu + \phi(h_t - \mu), 1)$. Using Proposition 3.2, the conditional distribution for κ_{t+1} is $[\kappa_{t+1}|h_t, \phi, \mu] \sim \text{TPB}(\beta, \alpha, \exp(\mu + \phi(h_t - \mu)))$. By substituting $\tau = \exp(\mu)$ and $\lambda_t = \exp(h_t - \mu)$, we equivalently have $[\kappa_{t+1}|\lambda_t, \phi, \tau] \sim \text{TPB}(\beta, \alpha, \tau^2\lambda_t^{2\phi})$. Noting $\tau^2\lambda_t^{2\phi} = \tau^{2(1-\phi)}\left[\frac{1-\kappa_t}{\kappa_t}\right]^{\phi}$ completes the proof. □

Proof. (Theorem 3.2) Let $\gamma_t = \big[(1-\kappa_t)/\kappa_t\big]^{\phi}$ and note that $\kappa \mapsto \kappa^{-1/2}$ and $\kappa \mapsto \big[1+(\gamma_t-1)\kappa\big]^{-1}$ are decreasing in κ for $\gamma_t > 1$. It follows that, for $\gamma_t > 1$,

$$\begin{aligned}
\mathbb{P}\big(\kappa_{t+1} > \varepsilon \big| \{\kappa_s\}_{s\leq t}, \phi\big) &= \int_{\varepsilon}^{1} \pi^{-1}\gamma_t^{1/2}\kappa_{t+1}^{-1/2}(1-\kappa_{t+1})^{-1/2}\left[1+(\gamma_t-1)\kappa_{t+1}\right]^{-1} d\kappa_{t+1} \\
&\leq \pi^{-1}\gamma_t^{1/2}\varepsilon^{-1/2}\left[1+(\gamma_t-1)\varepsilon\right]^{-1}\int_{\varepsilon}^{1}(1-\kappa_{t+1})^{-1/2}\,d\kappa_{t+1} \\
&\leq 2\pi^{-1}\varepsilon^{-1/2}(1-\varepsilon)^{1/2}\frac{\gamma_t^{1/2}}{1+(\gamma_t-1)\varepsilon}
\end{aligned}$$

converges to zero as $\kappa_t \to 0$, since $\kappa_t \to 0$ implies $\gamma_t \to \infty$. □

Proof. (Theorem 3.3) Marginalizing over ω_t, the likelihood is $[y_{t+1}|\{\kappa_s\}] \overset{indep}{\sim} N(0, \kappa_{t+1}^{-1})$. From Theorem 3.1, the posterior distribution of κ_{t+1} may be computed as

$$\begin{aligned}
[\kappa_{t+1}|y_{t+1}, \{\kappa_s\}_{s\leq t}, \phi, \tau] &\propto \left\{\kappa_{t+1}^{\beta-1}(1-\kappa_{t+1})^{\alpha-1}\big[1+(\gamma_t-1)\kappa_{t+1}\big]^{-(\alpha+\beta)}\right\} \\
&\quad \times \left\{\kappa_{t+1}^{1/2}\exp\left(-y_{t+1}^2\kappa_{t+1}/2\right)\right\} \\
&\propto (1-\kappa_{t+1})^{-1/2}\big[1+(\gamma_t-1)\kappa_{t+1}\big]^{-1}\exp\left(-y_{t+1}^2\kappa_{t+1}/2\right)
\end{aligned}$$

for $\alpha = \beta = 1/2$, where $\gamma_t = \tau^{2(1-\phi)}\big[(1-\kappa_t)/\kappa_t\big]^{\phi}$. Defining $p_1(\kappa) = (1-\kappa)^{-1/2}$, $p_2(\kappa|\gamma_t) = \big[1+(\gamma_t-1)\kappa\big]^{-1}$, and $p_3(\kappa|y_{t+1}) = \exp\left(-y_{t+1}^2\kappa/2\right)$ for $\kappa \in (0,1)$, observe that $p_1(\cdot)$ is increasing in κ, $p_2(\kappa|\gamma_t) \leq \big[p_1(\kappa)\big]^2$ for all $\gamma_t \geq 0$, and $p_3(\cdot)$ is decreasing in κ. Similar to Datta and Ghosh (2013), the following inequalities

hold for $\varepsilon \in (0,1)$ with $\varepsilon' = 1-\varepsilon$:

$$\begin{aligned}
\mathbb{P}\big(\kappa_{t+1} < \varepsilon' \big| y_{t+1}, \{\kappa_s\}_{s\le t}, \phi, \tau\big) &\le \frac{\mathbb{P}\big(\kappa_{t+1} < \varepsilon' \big| y_{t+1}, \{\kappa_s\}_{s\le t}, \phi, \tau\big)}{\mathbb{P}\big(\kappa_{t+1} > \varepsilon' \big| y_{t+1}, \{\kappa_s\}_{s\le t}, \phi, \tau\big)} \\
&\le \frac{\int_0^{\varepsilon'} (1-\kappa_{t+1})^{-3/2} \exp\left(-y_{t+1}^2 \kappa_{t+1}/2\right)\, d\kappa_{t+1}}{\int_{\varepsilon'}^1 \left[1+(\gamma_t-1)\kappa_{t+1}\right]^{-3/2} \exp\left(-y_{t+1}^2\kappa_{t+1}/2\right)\, d\kappa_{t+1}} \\
&\le \frac{\int_0^{\varepsilon'} (1-\kappa_{t+1})^{-3/2}\, d\kappa_{t+1}}{\exp\left(-y_{t+1}^2/2\right) \int_{\varepsilon'}^1 \left[1+(\gamma_t-1)\kappa_{t+1}\right]^{-3/2} d\kappa_{t+1}} \\
&\le \frac{2\big[(1-\varepsilon')^{-1/2}-1\big]}{\exp\left(-y_{t+1}^2/2\right) 2(\gamma_t-1)^{-1}\big\{\left[1+(\gamma_t-1)\varepsilon'\right]^{-1/2} - \gamma_t^{-1/2}\big\}} \\
&\le \big[(1-\varepsilon')^{-1/2}-1\big] \exp\left(y_{t+1}^2/2\right) \gamma_t^{1/2} \\
&\quad \times \left\{ \frac{1-\gamma_t}{1-\gamma_t^{1/2}/[1+(\gamma_t-1)\varepsilon']^{1/2}} \right\}.
\end{aligned}$$

Noting the final term in curly braces converges to 1 as $\gamma_t \to 0$, we obtain $\mathbb{P}\big(\kappa_{t+1} < 1-\varepsilon \big| y_{t+1}, \{\kappa_s\}_{s\le t}, \phi, \tau\big) \to 0$ as $\gamma_t \to 0$. The result for (a) follows immediately.

For $\varepsilon \in (0,1)$ and $\gamma_t < 1$, and observing that $p_2(\kappa|\gamma_t)$ is increasing in κ for $\gamma_t < 1$, then for any $\delta \in (0,1)$,

$$\begin{aligned}
\mathbb{P}\big(\kappa_{t+1} > \varepsilon \big| y_{t+1}, \{\kappa_s\}_{s\le t}, \phi, \tau\big) &\le \frac{\gamma_t^{-1} \exp\left(-y_{t+1}^2\varepsilon/2\right) \int_\varepsilon^1 (1-\kappa_{t+1})^{-1/2}\, d\kappa_{t+1}}{\int_0^{\delta\epsilon'} \exp\left(-y_{t+1}^2\kappa_{t+1}/2\right)\, d\kappa_{t+1}}, \\
&\le \frac{\gamma_t^{-1} \exp\left(-y_{t+1}^2\varepsilon/2\right) 2(1-\varepsilon)^{1/2}}{\exp\left(-y_{t+1}^2\delta\varepsilon/2\right)\delta\varepsilon} \\
&= \exp\left(-y_{t+1}^2\varepsilon[1-\delta]/2\right) \gamma_t^{-1} 2(1-\varepsilon)^{1/2}(\delta\varepsilon)^{-1}
\end{aligned}$$

which converges to zero as $|y_{t+1}| \to \infty$, proving (b). □

Proof. (Theorem 3.4) The density of $\eta \sim Z(\alpha,\beta,0,1)$ may be written

$$\begin{aligned}
[\eta] &= \frac{1}{B(\alpha,\beta)} \frac{[\exp(\eta)]^\alpha}{[1+\exp(\eta)]^{\alpha+\beta}} \\
&= \frac{1}{B(\alpha,\beta)} 2^{-(\alpha+\beta)} \exp\{\eta[\alpha-(\alpha+\beta)/2]\} \int_0^\infty \exp(-\eta^2\xi/2) p_{\alpha+\beta}(\xi)\, d\xi
\end{aligned}$$

using Theorem 1 of Polson et al. (2013), where $p_b(\xi)$ is the density of the random variable $\xi \sim \text{PG}(b, 0), b > 0$. It follows that

$$[\eta] \propto \int_0^\infty \exp\left\{-\frac{1}{2}\left[\eta^2\xi - \eta(\alpha-\beta)\right]\right\} p_{\alpha+\beta}(\xi)\, d\xi \propto \int_0^\infty f_N\big(\eta; \xi^{-1}[\alpha-\beta]/2, \xi^{-1}\big) p_{\alpha+\beta}(\xi)\, d\xi$$

where $f_N(\eta; \mu_N, \sigma_N^2)$ is the density of the random variable $\eta \sim N(\mu_N, \sigma_N^2)$.

The conditional distribution $[\xi|\eta] \sim \text{PG}(\alpha + \beta, \eta)$ is a direct result of Polson et al. (2013). □

B.1 MCMC Sampling Algorithm and Computational Details

We design a Gibbs sampling algorithm for the dynamic shrinkage process. The sampling algorithm is both computationally and MCMC efficient, and builds upon two main components: (1) a stochastic volatility sampling algorithm (Kastner and Frühwirth-Schnatter, 2014) augmented with a Pólya-Gamma sampler (Polson et al., 2013); and (2) a Cholesky Factor Algorithm (CFA, Rue, 2001) for sampling the state variables in the dynamic linear model. Alternative sampling algorithms exist for more general DLMs, such as the simulation smoothing algorithm of Durbin and Koopman (2002). However, as demonstrated by McCausland et al. (2011) and explored in Chan and Jeliazkov (2009) and Chan (2013), the CFA sampler is often more efficient. Importantly, both components employ algorithms that are linear in the number of time points, which produces a highly efficient sampling algorithm.

The general sampling algorithm is as follows, with the details provided in the subsequent sections:

1. Sample the dynamic shrinkage components (Section 3.5.1)

(a) Log-volatilities, $\{h_t\}$

(b) Pólya-Gamma mixing parameters, $\{\xi_t\}$

(c) Unconditional mean of log-volatility, μ

(d) AR(1) coefficient of log-volatility, ϕ

(e) Discrete mixture component indicators, $\{s_t\}$

2. Sample the state variables, $\{\boldsymbol{\beta}_t\}$ (Section B.1.2)

3. Sample the observation error variance, σ_ϵ^2.

For the observation error variance, we follow Carvalho et al. (2010) and assume the Jeffreys' prior $[\sigma_\epsilon^2] \propto 1/\sigma_\epsilon^2$. The full conditional distribution is $[\sigma_\epsilon|\{y_t\}_{t=1}^T, \{\beta_t\}_{t=1}^T, \tau^2] \propto \sigma_\epsilon^{-1} \times \sigma_\epsilon^{-T} \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\sum_{t=1}^T (y_t - \beta_t)^2\right\} \times \frac{\sqrt{T}}{\sigma_\epsilon(1+T\tau^2/\sigma_\epsilon^2)}$, where the last term arises from $\tau \sim C^+(0, \sigma_\epsilon/\sqrt{T})$. We sample from this distribution using the slice sampler of Neal (2003).

If we instead use a stochastic volatility model for the observation error variance as in Sections 3.3.2 and 3.4.2, we replace this step with a stochastic volatility sampling algorithm (e.g., Kastner and Frühwirth-Schnatter, 2014), which requires additional sampling steps for the corresponding log-volatility and the unconditional mean, AR(1) coefficient, and evolution error variance of log-volatility. An efficient implementation of such a sampler is available in the `R` package `stochvol` (Kastner, 2016). In this setting, we do not scale τ by the standard deviation, and instead assume $\tau \sim C^+(0, 1/\sqrt{T})$.

In Figure B.1, we provide empirical evidence for the linear time $\mathcal{O}(T)$ computations of the Bayesian trend filtering model with dynamic horseshoe innovations. The runtime per 1000 MCMC iterations is less than 6 minutes (on a

MacBook Pro, 2.7 GHz Intel Core i5) for samples sizes up to $T = 10^5$, so the Gibbs sampling algorithm is scalable.

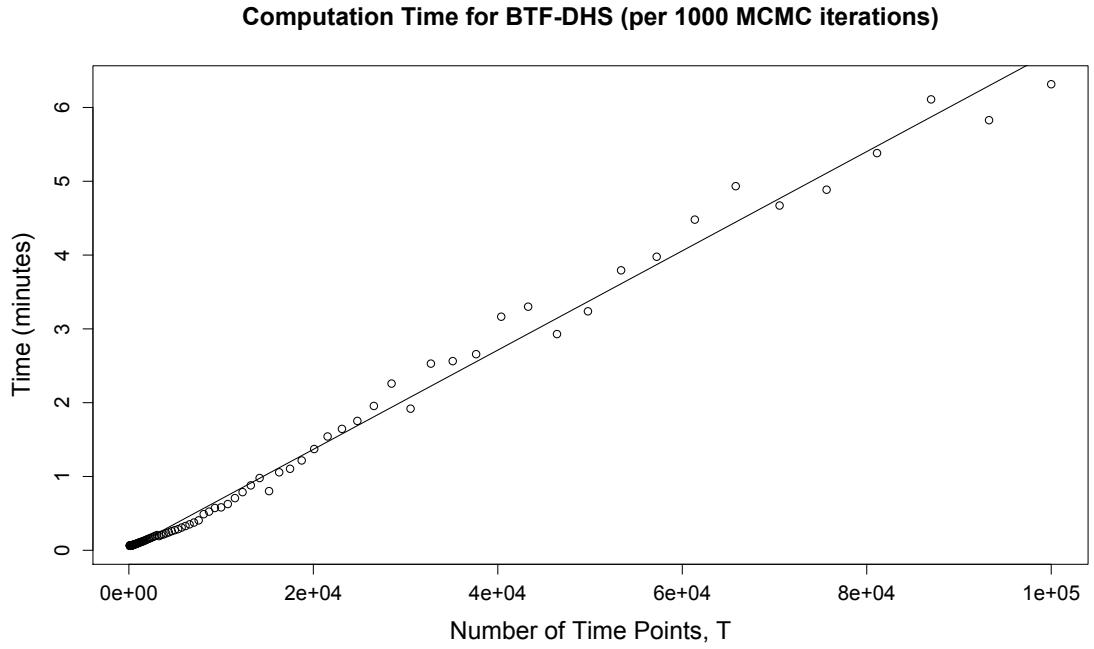


Figure B.1: Computation time per 1000 MCMC iterations for the Bayesian trend filtering model with dynamic horseshoe innovations (BTF-DHS).

B.1.1 Efficient Sampling for the Dynamic Shrinkage Process

Consider the (univariate) dynamic shrinkage process in (3.2) with the Pólya-Gamma parameter expansion of Theorem 3.4. We provide implementation details for the dynamic horseshoe prior with $\alpha = \beta = 1/2$, but extensions to other cases are straightforward. The SV sampling framework of Kastner and Frühwirth-Schnatter (2014) represents the likelihood for h_t on the log-scale, and approximates the ensuing $\log \chi_1^2$ distribution for the errors via a known discrete mixture of Gaussian distributions. In particular, let $\tilde{y}_t = \log(\omega_t^2 + c)$, where c

is a small offset to avoid numerical issues. Conditional on the mixture component indicators s_t, the likelihood is $\tilde{y}_t \overset{indep}{\sim} N(h_t + m_{s_t}, v_{s_t})$ where m_i and $v_i, i = 1, \ldots, 10$ are the pre-specified mean and variance components of the 10-component Gaussian mixture provided in Omori et al. (2007). The evolution equation is $h_{t+1} = \mu + \phi(h_t - \mu) + \eta_t$ with initialization $h_1 = \mu + \eta_0$ and innovations $[\eta_t|\xi_t] \overset{indep}{\sim} N(0, \xi_t^{-1})$ for $[\xi_t] \overset{iid}{\sim} \text{PG}(1, 0)$.

To sample $\boldsymbol{h} = (h_1, \ldots, h_T)$ jointly, we directly compute the posterior distribution of $\boldsymbol{h}$ and exploit the tridiagonal structure of the resulting posterior precision matrix. In particular, we equivalently have $\tilde{\boldsymbol{y}} \sim N(\boldsymbol{m} + \tilde{\boldsymbol{h}} + \tilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_v)$ and $\boldsymbol{D}_\phi \tilde{\boldsymbol{h}} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\xi)$, where $\boldsymbol{m} = (m_{s_1}, \ldots, m_{s_T})'$, $\tilde{\boldsymbol{h}} = (h_1 - \mu, \ldots, h_T - \mu)'$, $\tilde{\boldsymbol{\mu}} = (\mu, (1-\phi)\mu, \ldots, (1-\phi)\mu)'$, $\boldsymbol{\Sigma}_v = \text{diag}\left(\{v_{s_t}\}_{t=1}^T\right)$, $\boldsymbol{\Sigma}_\xi = \text{diag}\left(\{\xi_t^{-1}\}_{t=1}^T\right)$, and $\boldsymbol{D}_\phi$ is a lower triangular matrix with ones on the diagonal, $-\phi$ on the first off-diagonal, and zeros elsewhere. We sample from the posterior distribution of $\boldsymbol{h}$ by sampling from the posterior distribution of $\tilde{\boldsymbol{h}}$ and setting $\boldsymbol{h} = \tilde{\boldsymbol{h}} + \mu\boldsymbol{1}$ for $\boldsymbol{1}$ a T-dimensional vector of ones. The required posterior distribution is $\tilde{\boldsymbol{h}} \sim N\left(\boldsymbol{Q}_{\tilde{h}}^{-1}\boldsymbol{\ell}_{\tilde{h}}, \boldsymbol{Q}_{\tilde{h}}^{-1}\right)$, where $\boldsymbol{Q}_{\tilde{h}} = \boldsymbol{\Sigma}_v^{-1} + \boldsymbol{D}_\phi'\boldsymbol{\Sigma}_\xi^{-1}\boldsymbol{D}_\phi$ is a tridiagonal symmetric matrix with diagonal elements $\boldsymbol{d}_0(\boldsymbol{Q}_{\tilde{h}})$ and first off-diagonal elements $\boldsymbol{d}_1(\boldsymbol{Q}_{\tilde{h}})$ defined as

$$\boldsymbol{d}_0(\boldsymbol{Q}_{\tilde{h}}) = \Big[(v_{s_1}^{-1} + \xi_1 + \phi^2\xi_2), (v_{s_2}^{-1} + \xi_2 + \phi^2\xi_3), \ldots, (v_{s_{T-1}}^{-1} + \xi_{T-1} + \phi^2\xi_T), (v_{s_T}^{-1} + \xi_T)\Big],$$

$$\boldsymbol{d}_1(\boldsymbol{Q}_{\tilde{h}}) = \big[(-\phi\xi_2), (-\phi\xi_3), \ldots, (-\phi\xi_{T-1})\big], \text{ and}$$

$$\begin{aligned}\boldsymbol{\ell}_{\tilde{h}} &= \boldsymbol{\Sigma}_v^{-1}\left(\tilde{\boldsymbol{y}} - \boldsymbol{m} - \tilde{\boldsymbol{\mu}}\right) \\ &= \Big[\frac{\tilde{y}_1 - m_{s_1} - \mu}{v_{s_1}}, \frac{\tilde{y}_2 - m_{s_2} - (1-\phi)\mu}{v_{s_2}}, \ldots, \frac{\tilde{y}_T - m_{s_T} - (1-\phi)\mu}{v_{s_T}}\Big]'.\end{aligned}$$

Drawing from this posterior distribution is straightforward and efficient, using band back-substitution described in Kastner and Frühwirth-Schnatter (2014): (1) compute the Cholesky decomposition $\boldsymbol{Q}_{\tilde{h}} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is lower triangle;

(2) solve $\boldsymbol{L}\boldsymbol{a} = \boldsymbol{\ell}_{\tilde{h}}$ for $\boldsymbol{a}$; and (3) solve $\boldsymbol{L}'\tilde{\boldsymbol{h}} = \boldsymbol{a} + \boldsymbol{e}$ for $\tilde{\boldsymbol{h}}$, where $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}_T)$.

Conditional on the log-volatilities $\{h_t\}$, we sample the AR(1) evolution parameters: the log-innovation precisions $\{\xi_t\}$, the autoregressive coefficient ϕ, and the unconditional mean μ. The precisions are distributed $[\xi_t|\eta_t] \sim \text{PG}(1, \eta_t)$ for $\eta_t = h_{t+1} - \mu - \phi(h_t - \mu)$, which we sample using the `rpg()` function in the `R` package `BayesLogit` (Polson et al., 2013). The Pólya-Gamma sampler is efficient: using only exponential and inverse-Gaussian draws, Polson et al. (2013) construct an accept-reject sampler for which the probability of acceptance is uniformly bounded below at 0.99919, which does not require any tuning. Next, we assume the prior $[(\phi + 1)/2] \sim \text{Beta}(a_\phi, b_\phi)$, which restricts $|\phi| < 1$ for stationarity, and sample from the full conditional distribution of ϕ using the slice sampler of Neal (2003). We select $a_\phi = 10$ and $b_\phi = 2$, which places most of the mass for the density of ϕ in $(0, 1)$ with a prior mean of $2/3$ and a prior mode of $4/5$ to reflect the likely presence of persistent volatility clustering. The prior for the global scale parameter is $\tau \sim C^+(0, \sigma_\epsilon/\sqrt{T})$, which implies $\mu = \log(\tau^2)$ is $[\mu|\sigma_\epsilon, \xi_\mu] \sim N(\log(\sigma_\epsilon^2/T), \xi_\mu^{-1})$ with $\xi_\mu \sim \text{PG}(1, 0)$. Including the initialization $h_1 \sim N(\mu, \xi_0^{-1})$ with $\xi_0 \sim \text{PG}(1, 0)$, the posterior distribution for μ is $\mu \sim N(Q_\mu^{-1}\ell_\mu, Q_\mu^{-1})$ with $Q_\mu = \xi_\mu + \xi_0 + (1-\phi)^2 \sum_{t=1}^{T-1} \xi_t$ and $\ell_\mu = \xi_\mu \log(\sigma_\epsilon^2/T) + \xi_0 h_1 + (1-\phi) \sum_{t=1}^{T-1} \xi_t(h_{t+1} - \phi h_t)$. Sampling ξ_μ and ξ_0 follows the Pólya-Gamma sampling scheme above.

Finally, we sample the discrete mixture component indicators s_t. The discrete mixture probabilities are straightforward to compute: the prior mixture probabilities are the mixing proportions given by Omori et al. (2007) and the likelihood is $\tilde{y}_t \overset{indep}{\sim} N(h_t + m_{s_t}, v_{s_t})$; see Kastner and Frühwirth-Schnatter (2014) for details.

In the multivariate setting $p > 1$ of (3.10) with $\boldsymbol{\Phi} = \text{diag}\,(\phi_1, \ldots, \phi_p)$, we may modify the log-volatility sampler of $\{h_{j,t}\}$ by redefining relevant quantities using the ordering $\boldsymbol{h} = (h_{1,1}, \ldots, h_{1,T}, h_{2,1}, \ldots, h_{p,T})'$. In particular, the posterior precision matrix is again tridiagonal, but with diagonal elements $d_0(\boldsymbol{Q}_{\tilde{h}}) = \left[d_{0,1}(\boldsymbol{Q}_{\tilde{h}}), \ldots, d_{0,p}(\boldsymbol{Q}_{\tilde{h}})\right]$ and first off-diagonal elements $\boldsymbol{d}_1(\boldsymbol{Q}_{\tilde{h}}) = \left[d_{1,1}(\boldsymbol{Q}_{\tilde{h}}), 0, d_{1,2}(\boldsymbol{Q}_{\tilde{h}}), 0, \ldots, 0, d_{1,p}(\boldsymbol{Q}_{\tilde{h}})\right]$, where $d_{0,j}(\boldsymbol{Q}_{\tilde{h}})$ and $d_{1,j}(\boldsymbol{Q}_{\tilde{h}})$ are the diagonal elements and first off-diagonal elements, respectively, for predictor j as computed in the univariate case above. Similarly, the linear term $\boldsymbol{\ell}_{\tilde{h}} = \left[\boldsymbol{\ell}'_{\tilde{h},1}, \ldots, \boldsymbol{\ell}'_{\tilde{h},p}\right]'$ where $\boldsymbol{\ell}_{\tilde{h},j}$ is the linear term for predictor j as computed in the univariate case. The parameters $\xi_{j,t}$, ϕ_j, and $s_{j,t}$ may be sampled independently as in the univariate case, while samplers for $\{\mu_j\}$ and μ_0 proceed as in a standard hierarchical Gaussian model. For the more general case of non-diagonal $\boldsymbol{\Phi}$, we may use a simulation smoothing algorithm (e.g., Durbin and Koopman, 2002) for the log-volatilities $\{h_{j,t}\}$, while the sampler for $\boldsymbol{\Phi}$ will depend on the chosen prior.

B.1.2 Efficient Sampling for the State Variables

In the univariate setting of (3.8), the sampler for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_T)$ is similar to the log-volatility sample in Section 3.5.1. We provide the details for $D = 2$; the $D = 1$ case is similar to Section 3.5.1 with $\phi = 1$, $\mu = 0$, and $m_{s_t} = 0$. Model (3.8) may be written $\boldsymbol{y} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_\epsilon)$ and $\boldsymbol{D}_2\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\omega)$, where $\boldsymbol{y} = (y_1, \ldots, y_T)'$, $\boldsymbol{\Sigma}_\epsilon = \text{diag}\left(\{\sigma_t^2\}_{t=1}^T\right)$, $\boldsymbol{\Sigma}_\omega = \text{diag}\left(\{\sigma_{\omega_t}^2\}_{t=1}^T\right)$ for $\sigma_{\omega_t}^2 = \tau^2\lambda_t^2$, and $\boldsymbol{D}_2$ is a lower triangular matrix with ones on the diagonal, $(0, -2, \ldots, -2)$ on the first off-diagonal, ones on the second off-diagonal, and zeros elsewhere. Note that we allow the observation error variance σ_t^2 to be time-dependent for full gener-

ality, as in Section 3.4.2. The posterior for $\boldsymbol{\beta}$ is $\boldsymbol{\beta} \sim N\left(\boldsymbol{Q}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\ell}_{\boldsymbol{\beta}}, \boldsymbol{Q}_{\boldsymbol{\beta}}^{-1}\right)$, where $\boldsymbol{Q}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\epsilon}^{-1} + \boldsymbol{D}_2'\boldsymbol{\Sigma}_{\omega}^{-1}\boldsymbol{D}_2$ is a pentadiagonal symmetric matrix with diagonal elements $\boldsymbol{d}_0(\boldsymbol{Q}_{\boldsymbol{\beta}})$, first off-diagonal elements $\boldsymbol{d}_1(\boldsymbol{Q}_{\boldsymbol{\beta}})$, and second-off diagonal elements $\boldsymbol{d}_2(\boldsymbol{Q}_{\boldsymbol{\beta}})$ defined as

$$\begin{aligned}
\boldsymbol{d}_0(\boldsymbol{Q}_{\boldsymbol{\beta}}) = \Big[&\big(\sigma_1^{-2} + \sigma_{\omega_1}^{-2} + \sigma_{\omega_3}^{-2}\big), \big(\sigma_2^{-2} + \sigma_{\omega_2}^{-2} + 4\sigma_{\omega_3}^{-2} + \sigma_{\omega_4}^{-2}\big), \ldots, \\
&\big(\sigma_t^{-2} + \sigma_{\omega_t}^{-2} + 4\sigma_{\omega_{t+1}}^{-2} + \sigma_{\omega_{t+2}}^{-2}\big), \ldots, \\
&\big(\sigma_{T-2}^{-2} + \sigma_{\omega_{T-2}}^{-2} + 4\sigma_{\omega_{T-1}}^{-2} + \sigma_{\omega_T}^{-2}\big), \big(\sigma_{T-1}^{-2} + \sigma_{\omega_{T-1}}^{-2} + 4\sigma_{\omega_T}^{-2}\big), \big(\sigma_T^{-2} + \sigma_{\omega_T}^{-2}\big)\Big], \\
\boldsymbol{d}_1(\boldsymbol{Q}_{\boldsymbol{\beta}}) = [&-2\sigma_{\omega_3}^{-2}, -2\big(\sigma_{\omega_3}^{-2} + \sigma_{\omega_4}^{-2}\big), \ldots, -2\big(\sigma_{\omega_t}^{-2} + \sigma_{\omega_{t+1}}^{-2}\big), \ldots, -2\big(\sigma_{\omega_{T-1}}^{-2} + \sigma_{\omega_T}^{-2}\big), -2\sigma_{\omega_T}^{-2}], \\
\boldsymbol{d}_2(\boldsymbol{Q}_{\boldsymbol{\beta}}) = [&\sigma_{\omega_3}^{-2}, \ldots, \sigma_{\omega_t}^{-2}, \ldots, \sigma_{\omega_T}^{-2}],
\end{aligned}$$

and $\boldsymbol{\ell}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{y} = \left[y_1/\sigma_1^2, \ldots, y_t/\sigma_t^2, \ldots, y_T/\sigma_T^2\right]'$. Drawing from the posterior distribution is straightforward and efficient using the back-band substitution algorithm described in Section 3.5.1.

In the multivariate setting of (3.9), we similarly derive the posterior distribution for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_T')' = (\beta_{1,1}, \beta_{2,1}, \ldots, \beta_{p,1}, \beta_{1,2}, \ldots, \beta_{p,T})'$. Let $\boldsymbol{X} = \text{blockdiag}\left(\{\boldsymbol{x}_t'\}_{t=1}^T\right)$ denote the $T \times Tp$ block-diagonal matrix of predictors and $\boldsymbol{\Sigma}_{\omega} = \text{diag}\left(\{\sigma_{\omega_{j,t}}^2\}_{j,t}\right)$ for $\sigma_{\omega_{j,t}}^2 = \tau_0^2\tau_j^2\lambda_{j,t}^2$. The posterior distribution is $\boldsymbol{\beta} \sim N\left(\boldsymbol{Q}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\ell}_{\boldsymbol{\beta}}, \boldsymbol{Q}_{\boldsymbol{\beta}}^{-1}\right)$, where

$$\boldsymbol{Q}_{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{X} + \left(\boldsymbol{D}_2' \otimes \boldsymbol{I}_p\right)\boldsymbol{\Sigma}_{\omega}^{-1}\left(\boldsymbol{D}_2 \otimes \boldsymbol{I}_p\right)$$

and

$$\boldsymbol{\ell}_{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Sigma}_{\epsilon}^{-1}\boldsymbol{y} = \left[\boldsymbol{x}_1'y_1/\sigma_1^2, \ldots, \boldsymbol{x}_t'y_t/\sigma_t^2, \ldots, \boldsymbol{x}_T'y_T/\sigma_T^2\right]'.$$

Note that $\boldsymbol{Q}_{\boldsymbol{\beta}}$ may be constructed directly as above, but is now $2p$-banded. Alternatively, the regression coefficients $\{\beta_{j,t}\}$ may be sampled jointly using the simulation smoothing algorithm of Durbin and Koopman (2002).

B.2 Linear Regression for the Fama-French Asset Pricing Model

We present the ordinary linear regression results for the six-factor model discussed in Section 4.2, in which we append the momentum factor of Carhart (1997) to the five-factor Fama-French model (FF-5, Fama and French, 2015). We use weekly industry portfolio data from the website of Kenneth R. French, which provide the value-weighted return of stocks in the given industry. We focus on manufacturing (Manuf) and healthcare (Hlth). For a given industry portfolio, the response variable is the returns in excess of the risk free rate, $y_t = R_t - R_{F,t}$, with predictors $\boldsymbol{x}_t = (1, R_{M,t} - R_{F,t}, SMB_t, HML_t, RMW_t, CMA_t, MOM_t)'$, defined as follows: the *market risk factor*, $R_{M,t} - R_{F,t}$ is the return on the market portfolio $R_{M,t}$ in excess of the risk free rate $R_{F,t}$; the *size factor*, SMB_t (small minus big) is the difference in returns between portfolios of small and large market value stocks; the *value factor*, HML_t (high minus low) is the difference in returns between portfolios of high and low book-to-market value stocks; the *profitability factor*, RMW_t is the difference in returns between portfolios of robust and weak profitability stocks; the *investment factor*, CMA_t is the difference in returns between portfolios of stocks of low and high investment firms; and the *momentum factor*, MOM_t is the difference in returns between portfolios of stocks with high and low prior returns. These data are publicly available on Kenneth R. French's website, which provides additional details on the portfolios. We standardize all predictors and the response to have unit variance.

The results for the weekly manufacturing and healthcare industry data sets from 4/1/2007 - 4/1/2017 ($T = 522$) are in Tables B.1 and B.2, respectively. For

the manufacturing industry, the significant factors are market risk ($R_{M,t} - R_{F,t}$), profitability (RMW_t), and investment (CMA_t). For the healthcare industry, the significant factors are market risk, size (SMB_t), value (HML_t), and profitability.

Ordinary Linear Regression: Manufacturing Industry				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.020	0.015	-1.350	0.178
Mkt.RF	1.010	0.018	55.359	0.000
SMB	-0.013	0.016	-0.780	0.436
HML	-0.028	0.022	-1.264	0.207
RMW	0.088	0.018	4.918	0.000
CMA	0.052	0.017	3.106	0.002
MOM	0.029	0.020	1.437	0.151

Table B.1: Ordinary linear regression results for the weekly manufacturing industry data in the six-factor model. Significant factors at the 5% level are italicized.

Ordinary Linear Regression: Healthcare Industry				
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.045	0.023	1.966	0.050
Mkt.RF	0.924	0.028	32.686	0.000
SMB	-0.130	0.025	-5.221	0.000
HML	-0.264	0.034	-7.703	0.000
RMW	-0.168	0.028	-6.076	0.000
CMA	0.029	0.026	1.125	0.261
MOM	0.027	0.031	0.857	0.392

Table B.2: Ordinary linear regression results for the weekly healthcare industry data in the six-factor model. Significant factors at the 5% level are italicized.

APPENDIX C

FUNCTIONAL AUTOREGRESSION FOR SPARSELY SAMPLED DATA

C.1 Priors

The prior for $\{\boldsymbol{\mu}_t\}_{t=1}^T$ is determined by (4.6). Let $\boldsymbol{b}_\psi$ be a J_ψ-dimensional vector of cubic B-spline basis functions with $\min\{|\mathcal{T}_o|/2, 35\} = (J_\psi - 4)$ equally-spaced interior knots. The tensor product expansion $\psi_\ell(\tau, u) = \boldsymbol{b}_\psi'(\tau)\boldsymbol{\Theta}_{\psi_\ell}\boldsymbol{b}_\psi(u) = \left(\boldsymbol{b}_\psi'(u) \otimes \boldsymbol{b}_\psi'(\tau)\right)\boldsymbol{\theta}_{\psi_\ell}$, where $\boldsymbol{\Theta}_{\psi_\ell}$ is a $J_\psi \times J_\psi$ matrix of unknown coefficients and $\boldsymbol{\theta}_{\psi_\ell} = \text{vec}\,(\boldsymbol{\Theta}_{\psi_\ell})$, is computationally convenient for the FAR surfaces $\{\psi_\ell\}_{\ell=1}^p$. The Gaussian prior $[\boldsymbol{\theta}_{\psi_\ell}|\lambda_{\psi_\ell}] \sim N\left(\mathbf{0}, \lambda_{\psi_\ell}^{-1}\boldsymbol{\Omega}_{\psi_\ell}^{-1}\right)$ induces a Gaussian process prior on ψ_ℓ, where $\boldsymbol{\Omega}_{\psi_\ell}$ is a penalty matrix and λ_{ψ_ℓ} is a smoothing parameter. The standard roughness penalty $\int\int \left\{\frac{\partial^2}{\partial u_1}\psi_\ell(u_1, u_2) + 2\frac{\partial^2}{\partial u_1 \partial u_2}\psi_\ell(u_1, u_2) + \frac{\partial^2}{\partial u_2}\psi_\ell(u_1, u_2)\right\} du_1\, du_2$ can be expressed as $\boldsymbol{\theta}_{\psi_\ell}'\boldsymbol{\Omega}_2\boldsymbol{\theta}_{\psi_\ell}$ for a known singular matrix $\boldsymbol{\Omega}_2$. To obtain a proper prior, which is necessary for our model averaging procedure, we combine the roughness penalty with a nonstationarity penalty: a sufficient condition for stationarity of Y_t in model (4.2) is $\sum_{\ell=1}^p \int\int \psi_\ell^2(\tau, u)\, d\tau\, du < 1$, which can be expressed as $\sum_{\ell=1}^p \boldsymbol{\theta}_{\psi_\ell}'\boldsymbol{\Omega}_0\boldsymbol{\theta}_{\psi_\ell} < 1$ where $\boldsymbol{\Omega}_0$ is a known invertible matrix. We use the prior precision matrix $\boldsymbol{\Omega}_{\psi_\ell} = \boldsymbol{\Omega}_2 + \kappa_\ell\boldsymbol{\Omega}_0$, which penalizes roughness of ψ_ℓ and provides shrinkage toward stationarity, where the trade-off is determined by κ_ℓ. Simulations suggest that the posterior distribution is not sensitive to the choice of κ_ℓ; we fix $\kappa_\ell = 1$ for the simulations and assume $\log(\kappa_\ell) \sim N(0, 4)$ for the application. For the smoothing parameter λ_{ψ_ℓ}, we use the half-Cauchy prior of Gelman (2006), which provides excellent mixing of the states $\{s_\ell\}$ in the model averaging procedure. The prior may be expressed hierarchically via the auxil-

iary variables $\tilde{\lambda}_{\psi_\ell} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, $\tilde{\xi}_{\psi_\ell} \sim N\left(0, 10^6\right)$, and $\tilde{\boldsymbol{\theta}}_{\psi_\ell} \sim N\left(\mathbf{0}, \tilde{\lambda}_{\psi_\ell}^{-1}\boldsymbol{\Omega}_{\psi_\ell}^{-1}\right)$, with the identification $\boldsymbol{\theta}_{\psi_\ell} = \tilde{\xi}_{\psi_\ell}\tilde{\boldsymbol{\theta}}_{\psi_\ell}$ and $\lambda_{\psi_\ell} = \tilde{\xi}_{\psi_\ell}^{-2}\tilde{\lambda}_{\psi_\ell}$.

We use the conditionally conjugate inverse-Gamma priors $\sigma_\nu^{-2}, \sigma_\eta^{-2} \sim \text{Gamma}(10^{-3}, 10^{-3})$ for the measurement error precision and the FDLM approximation error precision, respectively. In some cases, we may prefer smoother sample paths of μ_t, but the paths will not be smooth when σ_η^2 is large. If increasing J_ϵ is infeasible or undesirable, fixing σ_η^2 at some small value, such as $\sigma_\eta^2 = 10^{-6}$, often works well, and can be interpreted as a jitter term for computing a valid inverse of $\boldsymbol{K}_\epsilon$ (Neal, 1999). Assuming the FDLM (4.7) for the innovation covariance $\boldsymbol{K}_\epsilon$, the factors are distributed $\boldsymbol{e}_t \overset{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_e)$ with $\boldsymbol{\Sigma}_e = \text{diag}\left(\{\sigma_j^2\}_{j=1}^{J_\epsilon}\right)$, although many generalizations are available (Kowal et al., 2016). To enforce the ordering constraints $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_{J_\epsilon}^2 > 0$, recall that the joint distribution (of the precisions) may be written $\left[\sigma_1^{-2}, \ldots, \sigma_{J_\epsilon}^{-2}\right] = \left[\sigma_{J_\epsilon}^{-2}\right] \prod_{j=1}^{J_\epsilon - 1} \left[\sigma_j^{-2} | \sigma_{j+1}^{-2}, \ldots, \sigma_{J_\epsilon}^{-2}\right]$. A noninformative joint prior that respects the constraints is fully specified by $\sigma_{J_\epsilon}^{-2} \sim \text{Gamma}\left(10^{-3}, 10^{-3}\right)$ and $\left[\sigma_j^{-2} | \sigma_{j+1}^{-2}, \ldots, \sigma_{J_\epsilon}^{-2}\right] = \left[\sigma_j^{-2} | \sigma_{j+1}^{-2}\right] \sim \text{Uniform}\left(0, \sigma_{j+1}^{-2}\right)$ for $j = 1, \ldots, J_\epsilon - 1$. The FLCs are $\phi_j(\tau) = \boldsymbol{b}_\phi'(\tau)\boldsymbol{\xi}_j$, where $\boldsymbol{b}_\phi$ is a low-rank thin plate spline basis with knot locations determined by the quantiles of the observation points, $\mathcal{T}_o$, $\boldsymbol{\xi}_j \sim N(\mathbf{0}, \boldsymbol{\Lambda}_{\phi_j})$, and $\boldsymbol{\Lambda}_{\phi_j}^{-1}$ is the low-rank thin plate spline penalty matrix. We follow Wand and Ormerod (2008) in the singular value decomposition-based diagonalization of the penalty matrix, so that $\boldsymbol{\Lambda}_\phi = \text{diag}\left(10^8, 10^8, \lambda_{\phi_j}^{-1}, \ldots, \lambda_{\phi_j}^{-1}\right)$, which places a noninformative prior on the constant and linear components of the thin plate spline basis, which are unpenalized. The prior precision λ_{ϕ_j} is common among the nonlinear components, and corresponds to the smoothing parameter for the regression function ϕ_j. Following Gelman (2006), we place uniform priors on the standard deviations $\lambda_{\phi_j}^{-1/2} \sim \text{Uniform}\left(0, 10^4\right)$, which im-

plies the prior for the precision $[\lambda_{\phi_j}] \propto \lambda_{\phi_j}^{-3/2}\mathbf{1}\{\lambda_{\phi_j} > 10^{-8}\}$. The upper bound for the prior standard deviation is selected to match the noninformative components of $\boldsymbol{\Lambda}_{\phi_j}$. The orthonormality constraint is enforced during sampling, which we discuss in Appendix C. We assume the same parametrization and prior distribution for the mean function, $\mu(\tau) = \boldsymbol{b}_\phi'(\tau)\boldsymbol{\theta}_\mu$.

C.2 Proof of Theorem 4.1

To prove Theorem 4.1, we use the following well-known results:

Proposition C.1. *For random vectors $\boldsymbol{\delta}$ and $\boldsymbol{Y}$ with known mean and covariance, the* unique best linear predictor *of $\boldsymbol{\delta}$ given $\boldsymbol{Y}$ is $\mathbb{E}_{\mathbb{G}}[\boldsymbol{\delta}|\boldsymbol{Y}]$, where $\mathbb{E}_{\mathbb{G}}$ is the expectation computed under the assumption that $(\boldsymbol{\delta}', \boldsymbol{Y}')'$ is jointly Gaussian.*

Proposition C.2 (West and Harrison, 1997)**.** *Under a DLM such as model* (4.6)*, the random vectors $\boldsymbol{y}_{1:T} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_T')'$ and $\boldsymbol{\mu}_{1:T} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_T')'$ are jointly Gaussian, conditional on the remaining parameters. In addition, all conditionals and marginals of the joint distribution of $(\boldsymbol{y}_{1:T}', \boldsymbol{\mu}_{1:T}')'$ are Gaussian.*

Note that we could extend $\boldsymbol{\mu}_{1:T}$ to include $\boldsymbol{\mu}$, which is also a Gaussian random vector. Following Propositions C.1 and C.2, the proof of Theorem 4.1 is straightforward:

Proof. (Theorem 4.1) Let $\mathcal{T}_e$ be fixed and finite such that $\mathcal{T}_e \subset \mathcal{T}$. Given this choice of $\mathcal{T}_e$, we can form the DLM (4.10) with the appropriately modified terms. Similarly, we can form the Gaussian DLM (4.6). Proposition C.2 implies that $(\boldsymbol{y}_{1:T}', \boldsymbol{\mu}_{1:T}')'$ under model (4.6) and conditional on $\boldsymbol{\Theta}$ is jointly Gaussian. Therefore, for any $\boldsymbol{\delta}, \boldsymbol{Y} \subseteq \mathcal{D}_T \cup \{\mu_t(\tau) : \tau \in \mathcal{T}_e, t = 1, \ldots, T\}$, i.e., any subvectors

of $(\boldsymbol{y}'_{1:T}, \boldsymbol{\mu}'_{1:T})'$, the distribution of $[\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$ is Gaussian. Proposition (C.1) implies that $\hat{\boldsymbol{\delta}}(\boldsymbol{Y}|\boldsymbol{\Theta}) \equiv \mathbb{E}[\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$, computed under the Gaussian DLM (4.6), is the unique best linear predictor of $[\boldsymbol{\delta}|\boldsymbol{Y}, \boldsymbol{\Theta}]$ under the DLM (4.10). □

C.3 Initialization and MCMC Sampling Algorithm

C.3.1 Initialization

We initialize the unknown functions using splines and the remaining parameters using conditional maximum likelihood estimators. We first estimate $\boldsymbol{\mu}$ as a smooth mean of $\{\boldsymbol{y}_t\}_{t=1}^T$, evaluated at $\mathcal{T}_e$. Next, we estimate each $\boldsymbol{\mu}_t$ by fitting a spline to $\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\mu}$ for $t = 1, \ldots, T$ using the `R` function `smooth.spline`. Since sparse observation points may lead to unstable initializations of $\boldsymbol{\mu}_t$, we compute the median degrees of freedom implied by the spline fits for $t = 1, \ldots, T$, and then recompute the splines for $t = 1, \ldots, T$ using this common degrees of freedom parameter. Conditional on these estimates, we estimate σ^2_ν, $\{\boldsymbol{\theta}_{\psi_1}, \ldots, \boldsymbol{\theta}_{\psi_p}\}$, and $\{\lambda_{\psi_1}, \ldots, \lambda_{\psi_p}\}$ using the maximum likelihood estimators, and initialize $\tilde{\boldsymbol{\theta}}_{\psi_\ell} = \boldsymbol{\theta}_{\psi_\ell}$, $\tilde{\lambda}_{\psi_\ell} = 1$, and $\tilde{\xi}_{\psi_\ell} = \lambda_{\psi_\ell}^{-1/2}$. From these estimators, we compute the innovations $\boldsymbol{\epsilon}_t$ for $t = 1, \ldots, T$. We initialize the FDLM parameters using the initialization algorithm of Kowal et al. (2016) based on the singular value decomposition (SVD) of $(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_T)' = \boldsymbol{U}_e\boldsymbol{D}_e\boldsymbol{V}'_e$. For the FLCs, we let $\boldsymbol{\Phi}$ equal the first J_ϵ columns of $\boldsymbol{V}_e$ and then estimate $\boldsymbol{\Xi}$ to minimize $||\boldsymbol{\Phi} - \boldsymbol{B}_\phi\boldsymbol{\Xi}||^2$. For the factors, we let $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_T)'$ be the first J_ϵ columns of $(\boldsymbol{U}_e\boldsymbol{D}_e)$, and then estimate $\{\sigma^2_j\}$ and σ^2_η using the conditional maximum likelihood estimators. Since $\sum_{k=1}^j \sigma^2_k / \sum_k \sigma^2_k$ estimates the proportion of variance of $\boldsymbol{\epsilon}_t$ explained by the first j

factors, we set J_ϵ to be the smallest number of factors that explain at least 95% of the variance of $\boldsymbol{\epsilon}_t$. While more sophisticated procedures are available for selecting J_ϵ, such as DIC and marginal likelihood, we find that this simple approach performs well in simulations.

C.3.2 Gibbs Sampling Algorithm

We propose to sample from the joint posterior distribution using a Gibbs sampler with the following steps:

1. FAR process, Y_t:

   (a) Centered FAR process, μ_t: form the DLM (6) and sample $\left[\{\boldsymbol{\mu}_t\}_{t=1}^T | \cdots\right]$ *jointly* using the state space sample of Durbin and Koopman (2002) implemented in the `R` package `KFAS`.

   (b) Mean function, $\mu(\tau) = \boldsymbol{b}_\phi'(\tau)\boldsymbol{\theta}_\mu$: sample $[\boldsymbol{\theta}_\mu | \cdots] \sim N(\boldsymbol{A}_\mu \boldsymbol{a}_\mu, \boldsymbol{A}_\mu)$ where

$$\boldsymbol{A}_\mu^{-1} = \boldsymbol{\Lambda}_\mu^{-1} + \sigma_\nu^{-2} \sum_{t=1}^T \boldsymbol{B}_\phi' \boldsymbol{Z}_t' \boldsymbol{Z}_t \boldsymbol{B}_\phi,$$

$$\boldsymbol{a}_\mu = \sigma_\nu^{-2} \sum_{t=1}^T \boldsymbol{B}_\phi' \boldsymbol{Z}_t' (\boldsymbol{y}_t - \boldsymbol{Z}_t \boldsymbol{\mu}_t),$$

   and $\boldsymbol{\Lambda}_\mu = \text{diag}\left(10^8, 10^8, \lambda_\mu^{-1}, \ldots, \lambda_\mu^{-1}\right)$. We sample the smoothing parameter $[\lambda_\mu | \cdots] \sim \text{Gamma}\left(\frac{1}{2}(J_\mu - 3), \frac{1}{2}\sum_{j=3}^{J_\mu} \theta_{\mu,j}^2\right)$ restricted to $\lambda_\mu > 10^{-8}$ (see the σ_j^{-2} sampler below), where J_μ $(= J_\phi)$ is the dimension of $\boldsymbol{\theta}_\mu$ and $\theta_{\mu,j}$ is the jth component of $\boldsymbol{\theta}_\mu$.

Set $Y_t = \mu_t + \mu$ or, in vector form, $\boldsymbol{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\mu}$.

2. Measurement error precision, σ_ν^{-2}: sample

$$[\sigma_\nu^{-2}|\cdots] \sim \text{Gamma}\left(10^{-3} + \frac{1}{2}\sum_{t=1}^{T} m_t, 10^{-3} + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{m_t}(y_{i,t} - \mu(\tau_{i,t}) - \mu_t(\tau_{i,t}))^2\right).$$

3. The FAR kernels, $\psi_1, \ldots, \psi_p$: using the Gelman (2006) prior and parametrization of $\boldsymbol{\theta}_{\psi_\ell} = \tilde{\xi}_{\psi_\ell}\tilde{\boldsymbol{\theta}}_{\psi_\ell}$, where $\psi_\ell(\tau, u) = \boldsymbol{b}'_\psi(\tau, u)\boldsymbol{\theta}_{\psi_\ell}$ and $\boldsymbol{B}_\psi = (\boldsymbol{b}_\psi(\tau_1), \ldots, \boldsymbol{b}_\psi(\tau_M))'$, we sample

   (a) $\tilde{\boldsymbol{\theta}}_\psi = (\tilde{\boldsymbol{\theta}}'_{\psi_1}, \ldots, \tilde{\boldsymbol{\theta}}'_{\psi_p})'$ *jointly* from $[\tilde{\boldsymbol{\theta}}_\psi|\cdots] \sim N(\boldsymbol{A}_\psi\boldsymbol{a}_\psi, \boldsymbol{A}_\psi)$, where

$$\boldsymbol{A}_\psi^{-1}[\ell,\ell] = \lambda_{\psi_\ell}\boldsymbol{\Omega}_{\psi_\ell} + s_\ell\tilde{\xi}_{\psi_\ell}^2\left[(\boldsymbol{B}'_\psi\boldsymbol{Q})\left\{\sum_{t=p+1}^{T}\boldsymbol{\mu}_{t-\ell}\boldsymbol{\mu}'_{t-\ell}\right\}(\boldsymbol{B}'_\psi\boldsymbol{Q})'\right] \otimes \left[\boldsymbol{B}'_\psi\boldsymbol{K}_\epsilon^{-1}\boldsymbol{B}_\psi\right],$$

$$\boldsymbol{A}_\psi^{-1}[\ell,k] = s_\ell s_k\tilde{\xi}_{\psi_\ell}\tilde{\xi}_{\psi_k}\left[(\boldsymbol{B}'_\psi\boldsymbol{Q})\left\{\sum_{t=p+1}^{T}\boldsymbol{\mu}_{t-\ell}\boldsymbol{\mu}'_{t-k}\right\}(\boldsymbol{B}'_\psi\boldsymbol{Q})'\right] \otimes \left[\boldsymbol{B}'_\psi\boldsymbol{K}_\epsilon^{-1}\boldsymbol{B}_\psi\right],$$

$$\boldsymbol{a}_\psi[\ell] = s_\ell\tilde{\xi}_{\psi_\ell}\text{vec}\left(\boldsymbol{B}'_\psi\boldsymbol{K}_\epsilon^{-1}\left\{\sum_{t=p+1}^{T}\boldsymbol{\mu}_t\boldsymbol{\mu}'_{t-\ell}\right\}(\boldsymbol{B}'_\psi\boldsymbol{Q})'\right),$$

   $\boldsymbol{A}_\psi^{-1}[\ell,k]$ is the (ℓ,k)th block of $\boldsymbol{A}_\psi^{-1}$ of dimension $J_\psi^2 \times J_\psi^2$ and $\boldsymbol{a}_\psi[\ell]$ is the ℓth subvector of $\boldsymbol{a}_\psi$ of length J_ψ^2;

   (b) For $\ell = 1, \ldots, p$, sample $\left[\tilde{\xi}_{\psi_\ell}|\cdots\right] \sim N\left(A_{\tilde{\xi}_{\psi_\ell}}a_{\tilde{\xi}_{\psi_\ell}}, A_{\tilde{\xi}_{\psi_\ell}}\right)$, where

$$A_{\tilde{\xi}_{\psi_\ell}}^{-1} = 10^{-6} + \tilde{\boldsymbol{\theta}}'_\psi\left(\left[(\boldsymbol{B}'_\psi\boldsymbol{Q})\left\{\sum_{t=p+1}^{T}\boldsymbol{\mu}_{t-\ell}\boldsymbol{\mu}'_{t-\ell}\right\}(\boldsymbol{B}'_\psi\boldsymbol{Q})'\right] \otimes \left[\boldsymbol{B}'_\psi\boldsymbol{K}_\epsilon^{-1}\boldsymbol{B}_\psi\right]\right)\tilde{\boldsymbol{\theta}}_\psi,$$

$$a_{\tilde{\xi}_{\psi_\ell}} = \tilde{\boldsymbol{\theta}}'_\psi\text{vec}\left(\boldsymbol{B}'_\psi\boldsymbol{K}_\epsilon^{-1}\left\{\sum_{t=p+1}^{T}\left[\boldsymbol{\mu}_t - \sum_{k\neq\ell}s_k\boldsymbol{G}(\psi_k)\boldsymbol{\mu}_{t-k}\right]\boldsymbol{\mu}'_{t-\ell}\right\}(\boldsymbol{B}'_\psi\boldsymbol{Q})'\right),$$

   sample $\left[\tilde{\lambda}_{\psi_\ell}|\cdots\right] \sim \text{Gamma}\left(\frac{1}{2} + J_\psi^2/2, \frac{1}{2} + \boldsymbol{\theta}'_{\psi_\ell}\boldsymbol{\Omega}_{\psi_\ell}\boldsymbol{\theta}_{\psi_\ell}/2\right)$, and, if κ_ℓ is unknown, sample κ_ℓ using the slice sampler (Neal, 2003). Set $\boldsymbol{\theta}_{\psi_\ell} = \tilde{\xi}_{\psi_\ell}\tilde{\boldsymbol{\theta}}_{\psi_\ell}$ and update $\boldsymbol{\Omega}_{\psi_\ell}$.

   (c) For the model averaging procedure, sample $[s_\ell|\cdots]$ (in random order), i.e., set $s_\ell = 1$ if $\log O_{10}^{post} > \log(1/U - 1)$ and $s_\ell = 0$ otherwise,

where $U \sim \text{Uniform}(0,1)$, $\log O_{10}^{post}$ is the log-posterior odds

$$\log O_{10}^{post} = -\frac{1}{2}\sum_{t=p+1}^{T}\left[\boldsymbol{\mu}_{t-\ell}'\boldsymbol{K}_{\epsilon}^{-1}\boldsymbol{\mu}_{t-\ell} - 2\Big(\boldsymbol{\mu}_t - \sum_{k\neq\ell} s_k \boldsymbol{G}(\psi_k)\boldsymbol{\mu}_{t-k}\Big)'\boldsymbol{K}_{\epsilon}^{-1}\boldsymbol{\mu}_{t-\ell}\right] + \log O_{10}^{prior},$$

and $\log O_{10}^{prior} = \log \mathbb{P}(s_\ell = 1 | s_k, k \neq \ell) - \log \mathbb{P}(s_\ell = 0 | s_k, k \neq \ell)$ is the log-prior odds.

4. The innovation covariance, $\boldsymbol{K}_\epsilon$, under the FDLM:

   (a) The factors, $\{\boldsymbol{e}_t\}_{t=1}^T$: using the prior $\boldsymbol{e}_t \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}_e)$ and the conditional likelihood $\boldsymbol{\epsilon}_t = \boldsymbol{\mu}_t - \sum_{\ell=1}^p \boldsymbol{G}(\psi_\ell)\boldsymbol{\mu}_{t-\ell} = \boldsymbol{\Phi}\boldsymbol{e}_t + \boldsymbol{\eta}_t$, sample $[\boldsymbol{e}_t | \cdots] \sim N(\boldsymbol{A}_e \boldsymbol{a}_{e_t}, \boldsymbol{A}_e)$, where

$$\boldsymbol{A}_e^{-1} = \sigma_\eta^{-2}\boldsymbol{\Phi}'\boldsymbol{\Phi} + \boldsymbol{\Sigma}_e^{-1} = \text{diag}\left(\{\sigma_\eta^{-2} + \sigma_j^{-2}\}_{j=1}^{J_\epsilon}\right)$$
$$\boldsymbol{a}_{e_t} = \sigma_\eta^{-2}\boldsymbol{\Phi}'\boldsymbol{\epsilon}_t.$$

   Note that $\boldsymbol{A}_e$ is time-invariant and diagonal, so we can sample $\{\boldsymbol{e}_t\}_{t=1}^T$ jointly and efficiently.

   (b) The factor precisions, σ_j^{-2}: sample

$$[\sigma_{J_\epsilon}^{-2} | \cdots] \sim \text{Gamma}\left(10^{-3} + \frac{T}{2}, 10^{-3} + \frac{1}{2}\sum_{t=1}^{T} e_{J_\epsilon,t}^2\right);$$

   then, for $j = J_\epsilon - 1, \ldots, 1$, set $\sigma_j^{-2} = F_\phi^{-1}(U; s_\phi, r_{\phi_j})$, where F_ϕ is the distribution function for a Gamma random variable with shape parameter $s_\phi = (T-1)/2$ and rate parameter $r_{\phi_j} = \sum_{t=1}^T e_{j,t}^2/2$, and $U \sim \text{Uniform}\left(a_{\phi_j}, b_{\phi_j}\right)$ where $a_{\phi_j} = F_\phi(0; s_\phi, r_{\phi_j})$ and $b_{\phi_j} = F_\phi(\sigma_{j+1}^{-2}; s_\phi, r_{\phi_j})$.

   (c) The approximation error precision, σ_η^{-2}: sample

$$[\sigma_\eta^{-2} | \cdots] \sim \text{Gamma}\left(10^{-3} + \frac{TM}{2}, 10^{-3} + \frac{1}{2}\sum_{t=1}^{T} ||\boldsymbol{\epsilon}_t - \boldsymbol{\Phi}\boldsymbol{e}_t||^2\right)$$

where $||\cdot||^2$ denotes the Euclidean distance.

(d) The factor loading curves: for $j = 1, \ldots, J_\epsilon$ (in random order), sample $\boldsymbol{\xi}_j \sim N(\boldsymbol{A}_{\xi_j}\boldsymbol{a}_{\xi_j}, \boldsymbol{A}_{\xi_j})$, where

$$\boldsymbol{A}_{\xi_j}^{-1} = \boldsymbol{\Lambda}_{\phi_j}^{-1} + \sigma_\eta^{-2}\left(\sum_{t=1}^{T} e_{j,t}^2\right)\boldsymbol{B}_\phi'\boldsymbol{B}_\phi,$$

$$\boldsymbol{a}_{\xi_j} = \sigma_\eta^{-2}\boldsymbol{B}_\phi'\sum_{t=1}^{T} e_{j,t}\left(\boldsymbol{\epsilon}_t - \boldsymbol{B}_\phi\sum_{k\neq j}\boldsymbol{\xi}_k e_{k,t}\right).$$

To enforce the orthogonality constraint, we condition on the linear constraints $(\boldsymbol{B}_\phi\boldsymbol{\xi}_k)'\boldsymbol{B}_\phi\boldsymbol{\xi}_j = 0$ for $k \neq j$; since $\boldsymbol{\xi}_j$ is Gaussian and $\boldsymbol{\xi}_k$ is conditioned upon, the resulting distribution is Gaussian with easily computable moments, which is also convenient for efficient sampling; see Kowal et al. (2016) for more details. After sampling from the conditional distribution, we normalize the sampled vector $\boldsymbol{\xi}_j$, so that the orthonormality constraint is enforced at every MCMC iteration. We sample the corresponding smoothing parameters $[\lambda_{\phi_j}|\cdots] \sim$ Gamma $\left(\frac{1}{2}(J_\phi - 3), \frac{1}{2}\sum_{k=3}^{J_\phi}\xi_{j,k}^2\right)$ restricted to $\lambda_{\phi_j} > 10^{-8}$, where $\xi_{j,k}$ is the kth component of $\boldsymbol{\xi}_j$.

Finally, we form the covariance and precision matrices $\boldsymbol{K}_\epsilon$ and $\boldsymbol{K}_\epsilon^{-1}$, respectively, using the sampled components. Since the orthonormality constraint $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_{J_\epsilon}$ is enforced at every MCMC iteration, we can compute $\boldsymbol{K}_\epsilon^{-1}$ directly and efficiently using (8).

When the sample size T or the number of evaluation points M is large (i.e., $T > 10,000$ or $M > 50$), the Durbin and Koopman (2002) joint sampler is computationally inefficient. Instead, we may use a single-move sampler for $\{\boldsymbol{\mu}_t\}_{t=1}^T$, in which we sample from the full conditional distribution of each $[\boldsymbol{\mu}_t|\boldsymbol{\mu}_s, s \neq t]$ separately for $t = 1, \ldots, T$ (in random order). The single-move sampler is more

computationally efficient, but is typically less MCMC efficient. The FDLM provides a closed form for $\boldsymbol{K}_{\epsilon}^{-1}$, which substantially reduces computation time when M is large.

The tensor product basis for ψ_ℓ provides a computational simplification for jointly sampling the FAR kernel basis coefficients, $\boldsymbol{\theta}_\psi$. Importantly, the dimension of the Kronecker product for computing $\boldsymbol{A}_{\psi}^{-1}$ is determined by the number of basis functions, J_ψ, which is bounded by 35 in our specification, and may be smaller for some applications. For other bivariate bases, such as the thin plate spline basis, such simplifications are not readily available, and the Kronecker product scales with the number of evaluation points, M.

In the model averaging procedure, there is a nontrivial concern about the ability of the MCMC sampler to move between states. When $s_\ell = 0$, ψ_ℓ does not appear in the likelihood (9), so the Gibbs sampler will draw ψ_ℓ from its prior. Therefore, the prior for ψ_ℓ must be proper; if it is nonetheless noninformative, then the draws of ψ_ℓ from the prior distribution may not be reasonable for (9), so the next MCMC sample of s_ℓ will be zero with high probability. To alleviate this problem, we fix $s_\ell = 1$ for all ℓ during a short burn-in period, so that each ψ_ℓ is well-estimated and therefore more likely to be included in the model if it is relevant. In both simulations and the yield curve application, the Gelman (2006) parametrization for ψ_ℓ sampling discussed in the Appendix provides excellent mixing among the states $\{s_\ell\}_{\ell=1}^{p_{max}}$.

C.4 Additional Theoretical Results

C.4.1 Proof of Proposition 4.1

Let $\Psi(B)$ be a polynomial in the backshift operator B of order p, so that $\Psi(B)Y_t = (1 - \Psi_1 B - \Psi_2 B^2 - \cdots - \Psi_p B^p)Y_t = Y_t - \sum_{\ell=1}^{p} \Psi_\ell(Y_{t-\ell})$, where $\{\Psi_\ell\}_{\ell=1}^{p}$ are bounded linear operators on $L^2(\mathcal{T})$. Similarly, let $\Theta(B)$ be a polynomial in the backshift operator B of order q, where $\{\Theta\}_{\ell=1}^{q}$ are bounded linear operators on $L^2(\mathcal{T})$. A *functional autoregressive moving average process* of order (p, q), written FARMA(p, q), is defined by $\Psi(B)(Y_t - \mu) = \Theta(B)\epsilon_t$, where $\{\epsilon_t\}$ is a white noise process in $L^2(\mathcal{T})$ and μ is the unconditional mean of Y_t. The FAR(p) model may be written compactly as $\Psi(B)(Y_t - \mu) = \epsilon_t$. By assumption, we observe the process $\{y_t\}$, where $y_t = Y_t + \nu_t$ and $\{\nu_t\}$ is a white noise process in $L^2(\mathcal{T})$ independent of $\{\epsilon_t\}$. Rewriting the observation equation $y_t - \mu = Y_t - \mu + \nu_t$ and applying $\Psi(B)$, we have $\Psi(B)(y_t - \mu) = \Psi(B)(Y_t - \mu) + \Psi(B)\nu_t = \epsilon_t + \Psi(B)\nu_t$. It remains to show that $Z_t \equiv \epsilon_t + \Psi(B)\nu_t$ is a *functional moving average process* of order p, or equivalently, FARMA($0, p$). Clearly, $X_t \equiv \Psi(B)\nu_t$ is FARMA($0, p$). By Proposition 10.2 in Bosq and Blanke (2008), $C_p^X \neq 0$ and $C_\ell^X = 0$ for $\ell > p$, where C_ℓ^X is the covariance operator of X_t defined by $C_\ell^X(x) \equiv \mathbb{E}\left[\langle X_t, x\rangle X_{t+\ell}\right]$ for $x \in L^2(\mathcal{T})$. Let C_ℓ^Z and C_ℓ^ϵ denote the covariance operators for Z_t and ϵ_t, respectively. Then $C_\ell^Z(x) = \mathbb{E}\left[\langle Z_t, x\rangle Z_{t+\ell}\right] = \mathbb{E}\left[\langle \epsilon_t + X_t, x\rangle \left(\epsilon_{t+\ell} + X_{t+\ell}\right)\right] = C_\ell^\epsilon(x) + C_\ell^X(x)$, using independence of $\{\epsilon_t\}$ and $\{\nu_t\}$. Since ϵ_t is white noise, $C_\ell^\epsilon = 0$ for $\ell > 0$, from which it follows that $C_p^Z \neq 0$ and $C_\ell^Z = 0$ for $\ell > p$. Proposition 10.2 in Bosq and Blanke (2008) implies that Z_t is FARMA($0, p$), so we conclude that y_t is FARMA(p, p).

C.4.2 DLM Recursions and Special Cases of Theorem 4.1

For completeness, we provide the standard DLM recursion formulas for model (6). Let $\mathcal{D}_t = \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1\} \cup \mathcal{D}_0$ be the information available at time t, where $\mathcal{D}_0$ represents the information prior to $t = 1$. For our purposes—in particular, for the Gibbs sampling algorithm—we let $\mathcal{D}_0 = \{\mu, \sigma_\nu^2, \psi, K_\epsilon\}$ (denoted by $\boldsymbol{\Theta}$ in Theorem 4.1). We may compute full conditional posterior distributions from model (6) using standard DLM recursions (e.g., West and Harrison, 1997). For simplicity, let $\boldsymbol{G} = \boldsymbol{G}(\psi)$. Suppose that $[\boldsymbol{\mu}_{t-1}|\mathcal{D}_{t-1}] \sim N(\boldsymbol{m}_{t-1}, \boldsymbol{C}_{t-1})$. The *prior at time t* is $[\boldsymbol{\mu}_t|\mathcal{D}_{t-1}] \sim N(\boldsymbol{a}_t, \boldsymbol{R}_t)$, where $\boldsymbol{a}_t = \boldsymbol{G}\boldsymbol{m}_{t-1}$ and $\boldsymbol{R}_t = \boldsymbol{G}\boldsymbol{C}_{t-1}\boldsymbol{G}' + \boldsymbol{K}_\epsilon$. The *one-step forecast at time t* is $[\boldsymbol{y}_t|\mathcal{D}_{t-1}] \sim N(\boldsymbol{f}_t, \boldsymbol{Q}_t)$, where $\boldsymbol{f}_t = \boldsymbol{Z}_t\boldsymbol{\mu} + \boldsymbol{Z}_t\boldsymbol{a}_t = \boldsymbol{Z}_t(\boldsymbol{\mu} + \boldsymbol{G}\boldsymbol{m}_{t-1})$ and $\boldsymbol{Q}_t = \boldsymbol{Z}_t\boldsymbol{R}_t\boldsymbol{Z}_t' + \sigma_\nu^2\boldsymbol{I}_{m_t}$. The *posterior at time t* is $[\boldsymbol{\mu}_t|\mathcal{D}_t] \sim N(\boldsymbol{m}_t, \boldsymbol{C}_t)$, where $\boldsymbol{m}_t = \boldsymbol{C}_t^{-1}\left(\boldsymbol{R}_t^{-1}\boldsymbol{a}_t + \sigma_\nu^{-2}\boldsymbol{Z}_t'(\boldsymbol{y}_t - \boldsymbol{Z}_t\boldsymbol{\mu})\right)$ and $\boldsymbol{C}_t^{-1} = \boldsymbol{R}_t^{-1} + \sigma_\nu^{-2}\boldsymbol{Z}_t'\boldsymbol{Z}_t$, or, more commonly, $\boldsymbol{m}_t = \boldsymbol{a}_t + \boldsymbol{A}_t\boldsymbol{r}_t$, $\boldsymbol{A}_t = \boldsymbol{R}_t\boldsymbol{Z}_t'\boldsymbol{Q}_t^{-1}$, $\boldsymbol{r}_t = \boldsymbol{y}_t - \boldsymbol{f}_t$, and $\boldsymbol{C}_t = \boldsymbol{R}_t - \boldsymbol{A}_t\boldsymbol{Q}_t\boldsymbol{A}_t'$. The h-step forecast of the functional observations is $\mathbb{E}[\boldsymbol{y}_{t+h}|\mathcal{D}_t] = \mathbb{E}[\boldsymbol{Z}_{t+h}\boldsymbol{\mu} + \boldsymbol{Z}_{t+h}\boldsymbol{\mu}_{t+h} + \boldsymbol{\nu}_{t+h}|\mathcal{D}_t] = \boldsymbol{Z}_{t+h}\boldsymbol{\mu} + \boldsymbol{Z}_{t+h}\mathbb{E}[\boldsymbol{\mu}_{t+h}|\mathcal{D}_t]$, where $\mathbb{E}[\boldsymbol{\mu}_{t+h}|\mathcal{D}_t] = \boldsymbol{G}^h\boldsymbol{m}_t$, which is the h-step forecast of $\boldsymbol{\mu}_t$.

Some special cases of Theorem 4.1 are proved in West and Harrison (1997):

Corollary B.2.1 (Theorem 4.10, West and Harrison, 1997). *The unique best linear predictor of the filtering random variable $[\boldsymbol{\mu}_t|\mathcal{D}_t]$ is $\boldsymbol{m}_t$*

Corollary B.2.2 (Corollary 4.7, West and Harrison, 1997). *The unique best linear predictor of the one-step forecast $[\boldsymbol{\mu}_t|\mathcal{D}_{t-1}]$ is $\boldsymbol{a}_t$. The unique best linear predictor of the one-step forecast $[\boldsymbol{y}_t|\mathcal{D}_{t-1}]$ is $\boldsymbol{f}_t$.*

C.4.3 Proof of Theorem 4.2

Suppose $\tau^* \in \mathcal{T}$ such that $\tau^* \notin \mathcal{T}_e$. The full conditional distribution of $\mu_t(\tau^*)$ is

$$\begin{aligned}\left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}, \mathcal{D}_s\right] &\propto \left[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s|\mu_t(\tau^*), \{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}\right] \times \left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}\right] \\ &\propto \left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}\right],\end{aligned}$$

since the likelihood term is constant with respect to $\mu_t(\tau^*)$: $\mathcal{T}_o \subseteq \mathcal{T}_e$, so $\tau^* \notin \mathcal{T}_e$ implies $\tau^* \notin \mathcal{T}_o$, and therefore $\mu_t(\tau^*)$ does not appear in the likelihood of model (4). For $p = 1$, the conditional Gaussian process prior for μ_t implied by model (4) under the approximation (5) is $[\mu_t|\mu_{t-1}, \psi, K_\epsilon] \sim \mathcal{GP}\left(\boldsymbol{\psi}'(\cdot)\boldsymbol{Q}\boldsymbol{\mu}_{t-1}, K_\epsilon\right)$, where $\boldsymbol{\psi}'(\tau) = (\psi(\tau, \tau_1), \ldots, \psi(\tau, \tau_M))$, $\boldsymbol{Q}$ is a known quadrature weight matrix, and $\boldsymbol{\mu}_{t-1} = (\mu_{t-1}(\tau_1), \ldots, \mu_{t-1}(\tau_M))'$ is the function μ_{t-1} evaluated at each $\tau \in \mathcal{T}_e$. Notably, $\tau^* \notin \mathcal{T}_e$ implies that $\mu_t(\tau^*)$ does not appear in the conditional mean function for μ_{t+1}, so we may further simplify the distribution of $\mu_t(\tau^*)$:

$$\left[\mu_t(\tau^*)|\{\boldsymbol{\mu}_r\}_{r=1}^T, \boldsymbol{\Theta}, \mathcal{D}_s\right] \propto \left[\mu_t(\tau^*)|\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}, \boldsymbol{\Theta}\right].$$

To compute this distribution, we use the definition of a Gaussian process, which implies the following joint distribution of $\mu_t(\tau^*)$ and $\boldsymbol{\mu}_t$, conditional on $\boldsymbol{\mu}_{t-1}$, ψ, and K_ϵ:

$$\begin{pmatrix} \mu_t(\tau^*) \\ \boldsymbol{\mu}_t \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} \\ \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1} \end{pmatrix}, \begin{pmatrix} K_\epsilon(\tau^*, \tau^*) & \boldsymbol{K}_\epsilon(\tau^*) \\ \boldsymbol{K}_\epsilon'(\tau^*) & \boldsymbol{K}_\epsilon \end{pmatrix}\right),$$

where $\boldsymbol{\Psi} = \{\psi(\tau_i, \tau_k)\}_{i,k=1}^M$ and $\boldsymbol{K}_\epsilon(\tau^*) = (K_\epsilon(\tau^*, \tau_1), \ldots, K_\epsilon(\tau^*, \tau_M))$. Conditioning on $\boldsymbol{\mu}_t$ induces the desired distribution $[\mu_t(\tau^*)|\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}, \psi, K_\epsilon] \sim N\left(m_t(\tau^*), K_t(\tau^*)\right)$, where $m_t(\tau^*) = \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{K}_\epsilon(\tau^*)\boldsymbol{K}_\epsilon^{-1}\left(\boldsymbol{\mu}_t - \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1}\right)$ and $K_t(\tau^*) = K_\epsilon(\tau^*, \tau^*) - \boldsymbol{K}_\epsilon(\tau^*)\boldsymbol{K}_\epsilon^{-1}\boldsymbol{K}_\epsilon'(\tau^*)$. Under the FDLM, the following useful simplifications are available: $K_\epsilon(\tau^*, \tau^*) = \sigma_\eta^2 + \boldsymbol{\phi}'(\tau^*)\boldsymbol{\Sigma}_e\boldsymbol{\phi}(\tau^*)$, $\boldsymbol{K}_\epsilon(\tau^*) = \boldsymbol{\phi}'(\tau^*)\boldsymbol{\Sigma}_e\boldsymbol{\Phi}'$, and using (8), $\boldsymbol{K}_\epsilon^{-1} = \sigma_\eta^{-2}\boldsymbol{I}_M - \sigma_\eta^{-2}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'$, where

$\boldsymbol{\phi}'(\tau^*) = (\phi_1(\tau^*), \ldots, \phi_{J_\epsilon}(\tau^*))$, $\boldsymbol{\Sigma}_e = \text{diag}\left(\{\sigma_j^2\}_{j=1}^{J_\epsilon}\right)$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}(\tau_1), \ldots, \boldsymbol{\phi}(\tau_M))'$, and $\tilde{\boldsymbol{\Sigma}}_e = \text{diag}\left(\{\sigma_j^2/(\sigma_\eta^2 + \sigma_j^2)\}_{j=1}^{J_\epsilon}\right)$. By substitution, we derive

$$\begin{aligned} m_t(\tau^*) &= \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{K}_\epsilon(\tau^*)\boldsymbol{K}_\epsilon^{-1}\left(\boldsymbol{\mu}_t - \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1}\right) \\ &= \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\phi}'(\tau^*)\boldsymbol{\Sigma}_e\boldsymbol{\Phi}'\left(\sigma_\eta^{-2}\boldsymbol{I}_M - \sigma_\eta^{-2}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'\right)\left(\boldsymbol{\mu}_t - \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1}\right) \\ &= \boldsymbol{\psi}'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-1} + \boldsymbol{\phi}'(\tau^*)\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'\left(\boldsymbol{\mu}_t - \boldsymbol{\Psi}\boldsymbol{Q}\boldsymbol{\mu}_{t-1}\right), \end{aligned}$$

using the constraint $\boldsymbol{\Phi}'\boldsymbol{\Phi} = \boldsymbol{I}_{J_\epsilon}$ and the simplification $\sigma_\eta^{-2}\boldsymbol{\Sigma}_e - \sigma_\eta^{-2}\boldsymbol{\Sigma}_e\tilde{\boldsymbol{\Sigma}}_e = \tilde{\boldsymbol{\Sigma}}_e$. Similarly,

$$\begin{aligned} K_t(\tau^*) &= K_\epsilon(\tau^*, \tau^*) - \boldsymbol{K}_\epsilon(\tau^*)\boldsymbol{K}_\epsilon^{-1}\boldsymbol{K}_\epsilon'(\tau^*) \\ &= \sigma_\eta^2 + \boldsymbol{\phi}'(\tau^*)\boldsymbol{\Sigma}_e\boldsymbol{\phi}(\tau^*) - \boldsymbol{\phi}'(\tau^*)\boldsymbol{\Sigma}_e\boldsymbol{\Phi}'\left(\sigma_\eta^{-2}\boldsymbol{I}_M - \sigma_\eta^{-2}\boldsymbol{\Phi}\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'\right)\boldsymbol{\Phi}\boldsymbol{\Sigma}_e\boldsymbol{\phi}(\tau^*) \\ &= \sigma_\eta^2 + \sigma_\eta^2\boldsymbol{\phi}'(\tau^*)\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\phi}(\tau^*), \end{aligned}$$

which is time-invariant. Extensions for $p > 1$ only require modification of the mean function: $m_t(\tau^*) = \sum_{\ell=1}^p \boldsymbol{\psi}_\ell'(\tau^*)\boldsymbol{Q}\boldsymbol{\mu}_{t-\ell} + \boldsymbol{\phi}'(\tau^*)\tilde{\boldsymbol{\Sigma}}_e\boldsymbol{\Phi}'\left(\boldsymbol{\mu}_t - \sum_{\ell=1}^p \boldsymbol{\Psi}_\ell\boldsymbol{Q}\boldsymbol{\mu}_{t-\ell}\right)$, where $\boldsymbol{\psi}_\ell'(\tau) = (\psi_\ell(\tau, \tau_1), \ldots, \psi_\ell(\tau, \tau_M))$ and $\boldsymbol{\Psi}_\ell = \{\psi_\ell(\tau_i, \tau_k)\}_{i,k=1}^M$.

C.5 Additional Simulation Results

In Figure C.1, we display the results from FAR(1) simulations under the dense design, while varying both smoothness of ϵ_t and the sample size, T. The functional data methods all nearly achieve the oracle performance, and are superior to the multivariate methods. These results confirm the findings of Didericksen et al. (2012): when T is large and the observation points are dense in $\mathcal{T}$, existing functional data methods can nearly achieve the oracle performance, even when ψ_1 is estimated poorly. The proposed methods, particularly with the FDLM (FDLM-FAR(1) and FDLM-FAR(p)), outperform existing functional

data methods for non-smooth GP innovations, and again are far superior for ψ_1 estimation. The uncertainty of p incorporated into the lag selection procedure (FDLM-FAR(p)) does not appear to inhibit forecasting or estimation of ψ_1 substantially.

For further clarity, we plot the *Bimodal-Gaussian* kernel in Figure C.2, which is featured prominently in our simulation study.

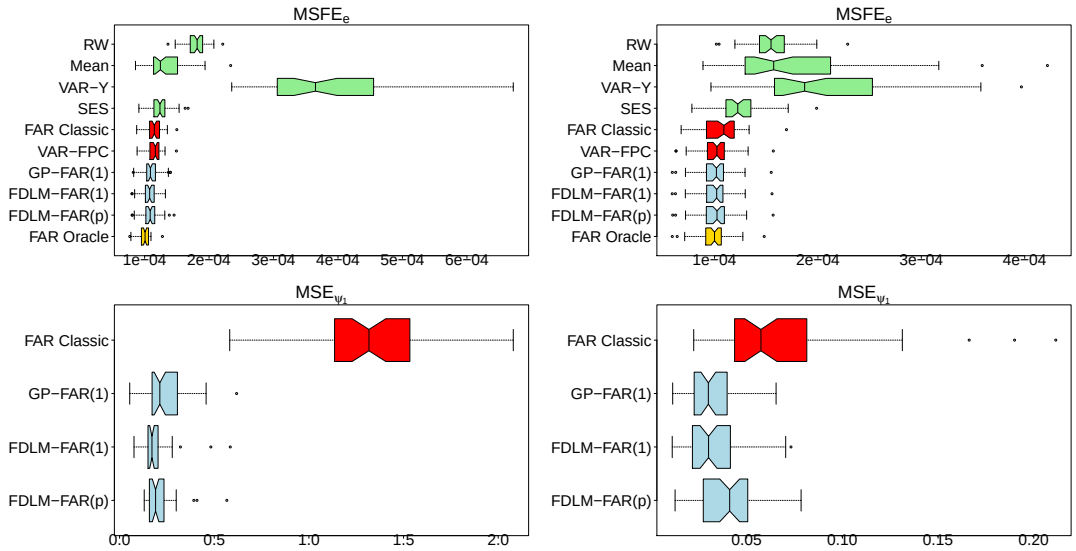


Figure C.1: $MSFE_e$ (**top**) and corresponding MSE_{ψ_1} (**bottom**) under various designs. **Left:** FAR(1), $T = 50$, dense design with the Bimodal-Gaussian kernel and non-smooth GP innovations. **Right:** FAR(1), $T = 350$, dense design with the Bimodal-Gaussian kernel and smooth GP innovations. The proposed methods provide superior forecasts and nearly achieve the oracle performance, despite the presence of sparsity.

C.6 Additional Details for the Yield Curve Application

We include MCMC diagnostics for the yield curve application. All diagnostics were computed using the `R` package `coda` (Plummer et al., 2006). In Figures C.3 and C.4, we provide trace plots for the one-step forecast distributions for

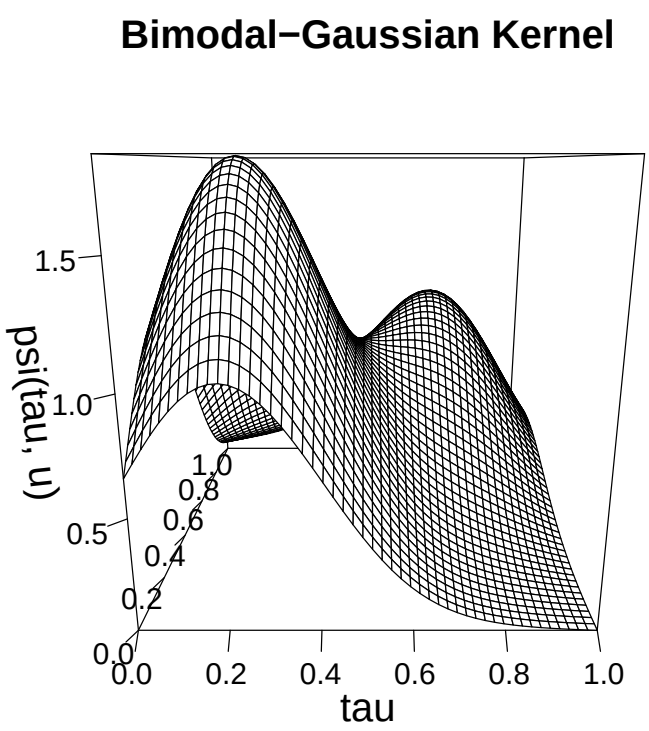


Figure C.2: The *Bimodal-Gaussian* kernel, $\psi(\tau, u) \propto \frac{0.75}{\pi(0.3)(0.4)} \exp\{-(\tau - 0.2)^2/(0.3)^2 - (u-0.3)^2/(0.4)^2\} + \frac{0.45}{\pi(0.3)(0.4)} \exp\{-(\tau-0.7)^2/(0.3)^2 - (u-0.8)^2/(0.4)^2\}$, normalized so that $\int \int \psi_\ell^2(\tau, u)\, d\tau\, du = 0.8$.

the nominal and real yield curves, respectively, on a single day in 2016 across selected maturities. The mixing is very efficient, which is confirmed by effective sample sizes which exceed 5,000 in all cases.

In our yield curve forecasting study of Section 4.7, we included two popular parametric yield curve models based on the Nelson-Siegel parametrization (Nelson and Siegel, 1987): Diebold and Li (2006, DL) and Diebold et al. (2006, DRA). The Nelson-Siegel basis is defined by $f_1(\tau) = 1$, $f_2(\tau|\lambda_{NS}) = \frac{1-\exp(-\tau\lambda_{NS})}{\tau\lambda_{NS}}$, and $f_3(\tau|\lambda_{NS}) = \frac{1-\exp(-\tau\lambda_{NS})}{\tau\lambda_{NS}} - \exp(-\tau\lambda_{NS})$, where λ_{NS} is an unknown parameter. For both DL and DRA, the yield curve $Y_t(\tau)$ for time t and time to maturity τ

is written as a linear combination of the Nelson-Siegel basis function, for which the corresponding weights are dynamic:

$$Y_t(\tau) = \boldsymbol{f}'(\tau|\lambda_{NS})\boldsymbol{\beta}_t + \epsilon_t(\tau), \tag{C.1}$$

$$(\boldsymbol{\beta}_t - \boldsymbol{\mu}_\beta) = \boldsymbol{A}\,(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}_\beta) + \boldsymbol{\eta}_t \tag{C.2}$$

where $\boldsymbol{f}'(\tau|\lambda_{NS}) = (f_1(\tau), f_2(\tau|\lambda_{NS}), f_3(\tau|\lambda_{NS}))$, $\boldsymbol{\beta}_t$ is the corresponding 3-dimensional vector of dynamic weights with unconditional mean $\boldsymbol{\mu}_\beta$, and $\boldsymbol{A}$ is the 3×3 evolution matrix. For implementation purposes, assume that the yield curve is observed at a fixed set of maturities $\tau_1, \ldots, \tau_M$, so that (C.1) becomes

$$\boldsymbol{y}_t = \boldsymbol{F}_{NS}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \tag{C.3}$$

where $\boldsymbol{y}_t = (Y_t(\tau_1), \ldots, Y_t(\tau_M))'$, $\boldsymbol{F}_{NS} = (\boldsymbol{f}(\tau_1|\lambda_{NS}), \ldots, \boldsymbol{f}(\tau_M|\lambda_{NS}))'$, and $\boldsymbol{\epsilon}_t = (\epsilon_t(\tau_1), \ldots, \epsilon_t(\tau_M))'$.

The DL approach fixes $\lambda_{NS} = 0.0609$ and then estimates the parameters using a multi-step procedure. First, the weights $\{\boldsymbol{\beta}_t\}$ are estimated using ordinary least squares from (C.3). Next, the evolution matrix $\boldsymbol{A}$ in (C.2) is estimated as a VAR coefficient matrix, conditional on $\{\boldsymbol{\beta}_t\}$. Diebold and Li (2006) note that constraining $\boldsymbol{A}$ to be diagonal may improve forecasting in some cases. Finally, h-step forecasts $\hat{\boldsymbol{y}}_{T+h}$ are computed via $\hat{\boldsymbol{y}}_{T+h} = \boldsymbol{F}_{NS}\hat{\boldsymbol{\beta}}_{T+h}$, where $\hat{\boldsymbol{\beta}}_{T+h}$ is the h-step forecast computed from the VAR in (C.2).

Alternatively, the DRA approach combines (C.3) and (C.2) into a state space model, with error distributions $\boldsymbol{\epsilon}_t \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{H})$ independent of $\boldsymbol{\eta}_t \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{Q})$. DRA assume that $\boldsymbol{H}$ is diagonal; we further assume that $\boldsymbol{Q}$ is diagonal, which helps stabilize computations. The unknown parameters $\{\lambda_{NS}, \boldsymbol{A}, \boldsymbol{H}, \boldsymbol{Q}\}$ are then estimated jointly using maximum likelihood based on the Kalman filter. Following DRA, we model λ_{NS} and the diagonal elements of $\boldsymbol{H}$ and $\boldsymbol{Q}$ on the

log-scale to ensure positivity in the optimization routine. Conditional on the maximum likelihood estimates for these parameters, DRA use standard state space computations to construct forecasts for the response vector, $\boldsymbol{y}_t$.

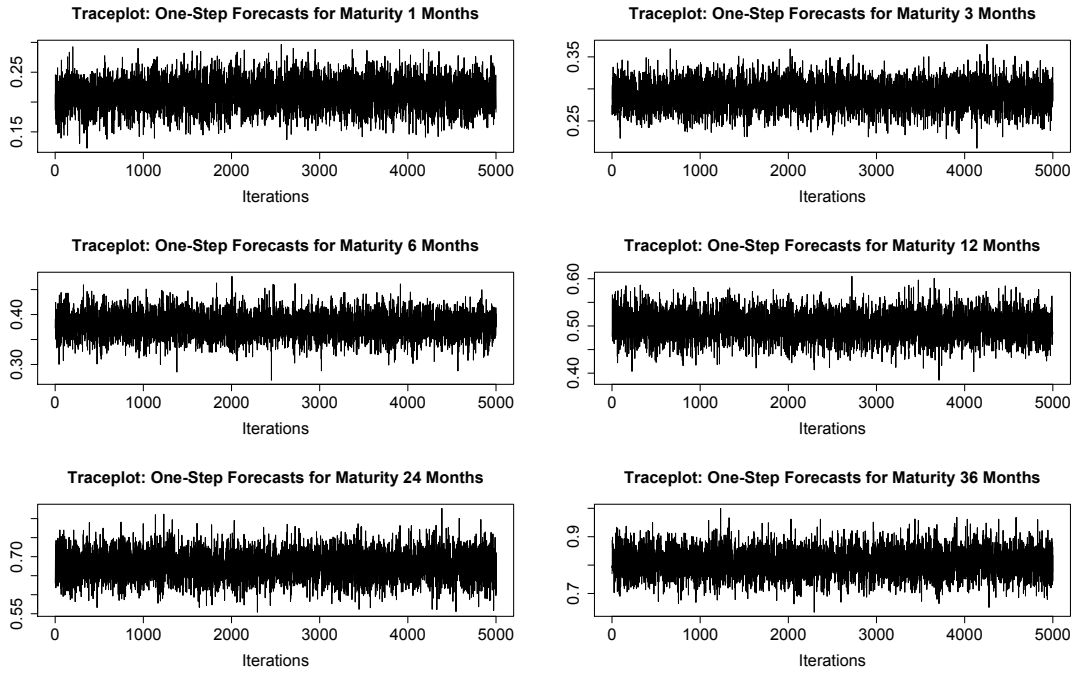


Figure C.3: Traceplot for one-step forecasts for nominal yield curves at selected maturities during 2016.

C.7 Additional Details on the Quadrature Approximation

Consider the integral in the FAR(1) evolution equation, $\mathcal{I}(\tau) \equiv \int \psi(\tau, u)\mu_{t-1}(u)\, du$, where we omit dependence of $\mathcal{I}$ on t for notational simplicity. In the proposed methodology, we approximate this integral using quadrature: $\mathcal{I}(\tau) \approx \mathcal{I}_M(\tau) \equiv (\psi(\tau, \tau_1), \ldots, \psi(\tau, \tau_M))\, \boldsymbol{Q}\boldsymbol{\mu}_{t-1}$, where $\{\tau_1, \ldots, \tau_M\} = \mathcal{T}_e \subset \mathcal{T}$ is the set of unique evaluation points, $\boldsymbol{Q}$ is a known $M \times M$ quadrature matrix, and $\boldsymbol{\mu}_{t-1} = (\mu_{t-1}(\tau_1), \ldots, \mu_{t-1}(\tau_M))'$ is the function μ_{t-1} evaluated at the evaluation

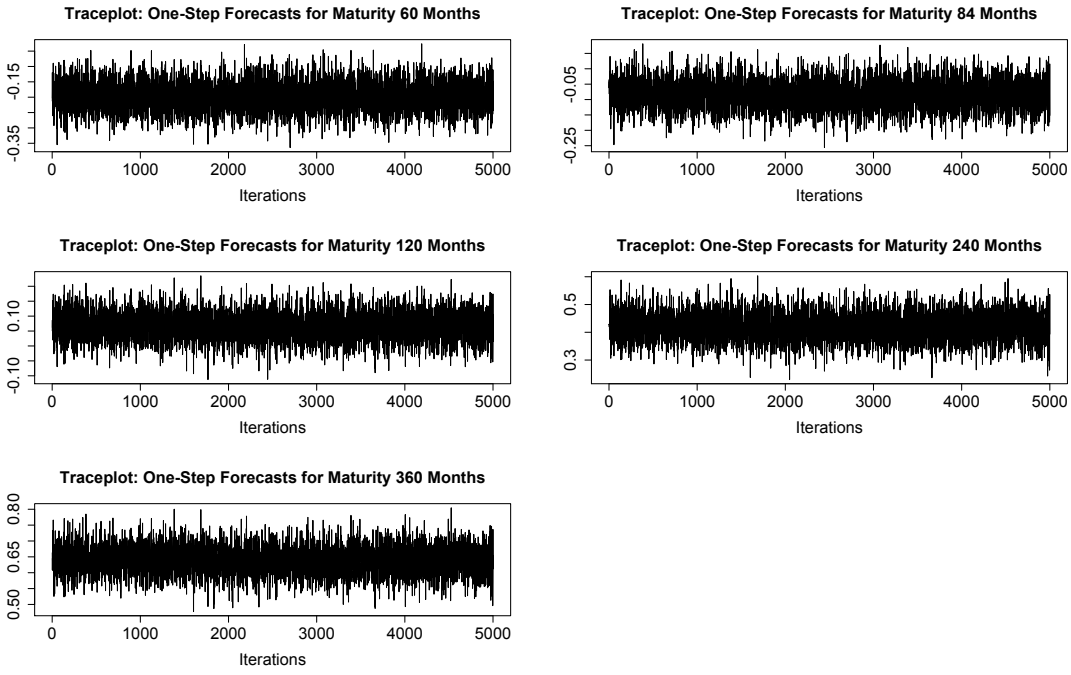


Figure C.4: Traceplot for one-step forecasts for real yield curves at selected maturities during 2016.

points. It is important to assess how the accuracy of the approximation of $\mathcal{I}$ by $\mathcal{I}_M$ depends in M, and in particular to determine a value of M sufficiently large to produce reasonable approximations in practice. However, there is a tradeoff: the state vector in the dynamic linear model is M-dimensional, so increasing M indiscriminately may unnecessarily increase computation time.

We conducted a sensitivity analysis based on the simulations from Section 4.6 of the main paper. In particular, we use the Bimodal-Gaussian kernel, $\psi(\tau, u) \propto \frac{0.75}{\pi(0.3)(0.4)} \exp\{-(\tau-0.2)^2/(0.3)^2-(u-0.3)^2/(0.4)^2\}+\frac{0.45}{\pi(0.3)(0.4)} \exp\{-(\tau-0.7)^2/(0.3)^2 - (u - 0.8)^2/(0.4)^2\}$, normalized so that $\int\int \psi_\ell^2(\tau, u)\, d\tau\, du = 0.8$. The Bimodal-Gaussian kernel is nonlinear, and therefore is inherently more difficult to approximate using linear quadrature methods, such as the trapezoidal rule. For the other component of the integrand, μ_{t-1},

we simulate $\mu_{t-1} \sim \mathcal{GP}(0, K_\epsilon)$ using the covariance function parameterization $K_\epsilon = \sigma^2 R_\rho$, where R_ρ is the Matérn correlation function $R_\rho(\tau, u) = \{2^{\rho_1 - 1}\Gamma(\rho_1)\}^{-1} (||\tau - u||/\rho_2)^{\rho_1} K_{\rho_1}(||\tau - u||/\rho_2)$, $\Gamma(\cdot)$ is the gamma function, K_{ρ_1} is the modified Bessel function of order ρ_1, and $\boldsymbol{\rho} = (\rho_1, \rho_2)$ are parameters (Matérn, 2013). We let $\sigma = 0.01$ and $\boldsymbol{\rho} = (\rho_1, 0.1)$, with $\rho_1 = 2.5$ for smooth (twice-differentiable) sample paths and $\rho_1 = 0.5$ for non-smooth (continuous, non-differentiable) sample paths. Comparisons between these cases are important: the non-smooth setting is substantially more challenging for approximations.

For each simulated value of $\mu_{t-1} \sim \mathcal{GP}(0, K_\epsilon)$, we compute $\mathcal{I}_{200}(\tau)$, which we use as a proxy for the true (but unknown) integral value $\mathcal{I}(\tau)$, and compare it to $\mathcal{I}_M(\tau)$ for $M \in \{5, 10, 15, 20, 25, 30, 40, 50, 60, 70, 80, 90, 100\}$. Note that the approximation induced by $\mathcal{I}_{200}(\tau)$ is also used to generate the simulations of Section 4.6 in the main paper. We measure accuracy using the *relative absolute error* (RAE) and the *standardized squared error* (SSE), defined respectively by

$$R_M = \int \left| \frac{\mathcal{I}_{200}(\tau) - \mathcal{I}_M(\tau)}{\mathcal{I}_{200}(\tau)} \right| d\tau, \quad S_M = \int \frac{(\mathcal{I}_{200}(\tau) - \mathcal{I}_M(\tau))^2}{\sigma^2} d\tau, \tag{C.4}$$

which we compute for each simulation. We report the pointwise medians for each R_M and S_M as a function of M in Figure E1. As expected, for fixed M, the integral approximation is more accurate when μ_{t-1}—and therefore the integrand—is smooth. Nonetheless, the relative gains of increasing M decline quickly for $M > 20$ in both cases.

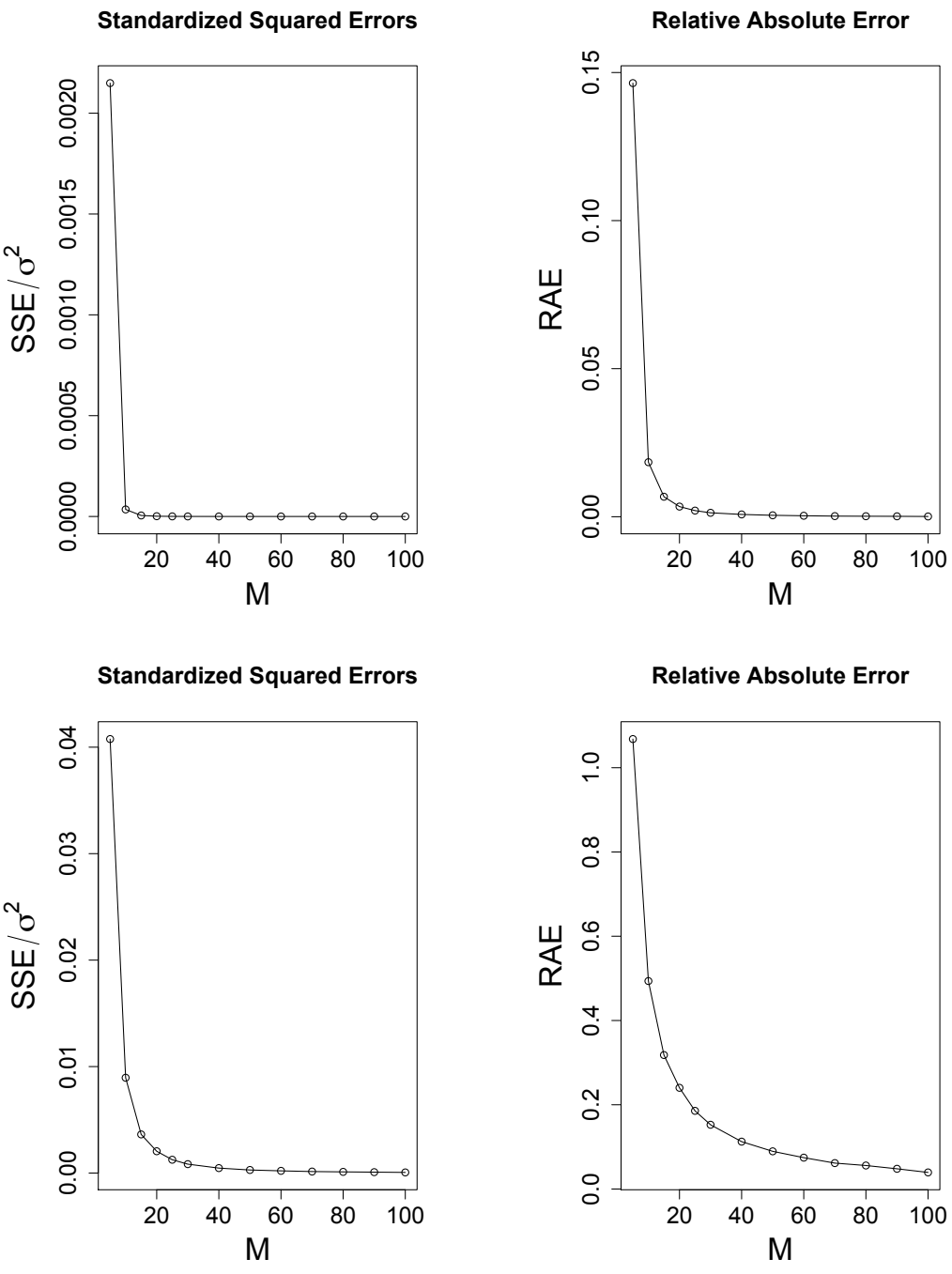


Figure E1: Standardized squared errors and relative absolute errors for smooth (**top**) and non-smooth (**bottom**) integrands. The errors are small in magnitude, particularly in the smooth case, and decay quickly for $M > 20$.

BIBLIOGRAPHY

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749.

Albert, J. H. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business & Economic Statistics*, 11(1):1–15.

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized Beta mixtures of Gaussians. In *Advances in neural information processing systems*, pages 523–531.

Arnold, T. B. and Tibshirani, R. J. (2014). `genlasso`*: Path algorithm for generalized lasso problems*. `R` package version 1.3.

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110(509):378–392.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–3430.

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 145–159.

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, 92(2):419–434.

Belmonte, M. A., Koop, G., and Korobilis, D. (2014). Hierarchical shrinkage in time-varying parameter models. *Journal of Forecasting*, 33(1):80–94.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, pages 716–761.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

Besse, P. C., Cardot, H., and Stephenson, D. B. (2000). Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, pages 673–687.

Bloomfield, P. (2004). *Fourier analysis of time series*. John Wiley & Sons.

Bolder, D., Johnson, G., and Metzler, A. (2004). *An empirical analysis of the Canadian term structure of zero-coupon interest rates*. Bank of Canada.

Bosq, D. (2000). *Linear processes in function spaces: theory and applications*, volume 149. Springer Science & Business Media.

Bosq, D. and Blanke, D. (2008). *Inference and prediction in large dimensions*, volume 754. John Wiley & Sons.

Botly, L. C. and De Rosa, E. (2009). Cholinergic deafferentation of the neocortex using 192 IgG-saporin impairs feature binding in rats. *The Journal of Neuroscience*, 29(13):4120–4130.

Bowsher, C. G. and Meeks, R. (2008). The dynamics of economic functions: modeling and forecasting the yield curve. *Journal of the American Statistical Association*, 103(484).

Cardot, H., Ferraty, F., and Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, 45(1):11–22.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Carvalho, C. M., Lopes, H. F., and Aguilar, O. (2011). Dynamic stock selection strategies: A structured factor model framework. *Bayesian Statistics*, 9:1–21.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *AISTATS*, volume 5, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, pages 465–480.

Chan, J. C. (2013). Moving average stochastic volatility models with application to inflation forecast. *Journal of Econometrics*, 176(2):162–172.

Chan, J. C. and Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1-2):101–120.

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367.

Chen, Y. and Li, B. (2015). An adaptive functional autoregressive forecast model to predict electricity price curves. *Journal of Business & Economic Statistics*, (just-accepted):1–56.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Ergashev, B. (2009). Analysis of multifactor affine yield curve models. *Journal of the American Statistical Association*, 104(488):1324–1337.

Chib, S., Nardari, F., and Shephard, N. (2002). Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108(2):281–316.

Constantine, W. and Percival, D. (2016). `wmtsa`: *Wavelet Methods for Time Series Analysis*. `R` package version 2.0-2.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of Statistical Software*, 14(14):1–24.

Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. John Wiley & Sons.

Cruz-Marcelo, A., Ensor, K. B., and Rosner, G. L. (2011). Estimating the term structure with a semiparametric Bayesian hierarchical model: an application to corporate bonds. *Journal of the American Statistical Association*, 106(494).

Damon, J. and Guillas, S. (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics*, 13:759–774.

Damon, J. and Guillas, S. (2005). Estimation and simulation of autoregressive hilbertian processes with exogenous variables. *Statistical Inference for Stochastic Processes*, 8(2):185–204.

Daníelsson, J. (1998). Multivariate stochastic volatility models: estimation and a comparison with VGARCH models. *Journal of Empirical Finance*, 5(2):155–173.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

Dées, S. and Saint-Guilhem, A. (2011). The role of the united states in the global economy and its evolution over time. *Empirical Economics*, 41(3):573–591.

Didericksen, D., Kokoszka, P., and Zhang, X. (2012). Empirical properties of forecasts with the functional autoregressive model. *Computational Statistics*, 27(2):285–298.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.

Diebold, F. X., Li, C., and Yue, V. Z. (2008). Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach. *Journal of Econometrics*, 146(2):351–363.

Diebold, F. X., Rudebusch, G. D., and Aruoba, B. S. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics*, 131(1):309–338.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616.

Earls, C. and Hooker, G. (2014). Bayesian covariance estimation and inference in latent Gaussian process models. *Statistical Methodology*, 18:79–100.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC Press.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Faulkner, J. R. and Minin, V. N. (2016). Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*.

Fernandez, C. and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16(01):80–101.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer.

Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–1159.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85–100.

Gamerman, D. and Migon, H. S. (1993). Dynamic hierarchical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 629–642.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.

Griffin, J. E. and Brown, P. J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, Centre for Research in Statistical Methodology.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Gu, C. (1992). Penalized likelihood regression: a Bayesian analysis. *Statistica Sinica*, 2(1):255–264.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

Hays, S., Shen, H., and Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, 6(3):870–894.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*, volume 200. Springer Science & Business Media.

Hyndman, R. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, 27(1):1–22.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

James, N. A., Kejariwal, A., and Matteson, D. S. (2016). Leveraging cloud data

to mitigate user experience from 'Breaking Bad'. In *2016 IEEE International Conference on Big Data*, pages 3499–3508. IEEE.

Jungbacker, B., Koopman, S. J., and van der Wel, M. (2013). Smooth dynamic factor analysis with application to the US term structure of interest rates. *Journal of Applied Econometrics*.

Kalli, M. and Griffin, J. E. (2014). Time-varying sparsity in dynamic regression models. *Journal of Econometrics*, 178(2):779–793.

Kargin, V. and Onatski, A. (2008). Curve forecasting by functional autoregression. *Journal of Multivariate Analysis*, 99(10):2508–2526.

Kastner, G. (2016). Dealing with stochastic volatility in time series using the `R` package `stochvol`. *Journal of Statistical Software*, 69(5):1–30.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kaufman, C. G. and Sain, S. R. (2010). Bayesian functional ANOVA modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3):361–393.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). ℓ_1 Trend Filtering. *SIAM review*, 51(2):339–360.

Kokoszka, P. (2012). Dependent functional data. *ISRN Probability and Statistics*, 2012.

Kokoszka, P. and Reimherr, M. (2013). Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34(1):116–129.

Koopman, S. J. and Durbin, J. (2000). Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis*, 21(3):281–296.

Koopman, S. J. and Durbin, J. (2003). Filtering and smoothing of state vector for diffuse state-space models. *Journal of Time Series Analysis*, 24(1):85–98.

Koopman, S. J., Mallee, M. I., and Van der Wel, M. (2010). Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters. *Journal of Business & Economic Statistics*, 28(3):329–343.

Korobilis, D. (2013a). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, 29(1):43–59.

Korobilis, D. (2013b). VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230.

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2016). A Bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association*. (in press).

Kowal, D. R., Matteson, D. S., and Ruppert, D. (2017). Functional autoregression for sparsely sampled data. *Journal of Business & Economic Statistics*, pages 1–13.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Laurini, M. P. (2014). Dynamic functional data analysis with non-parametric state space models. *Journal of Applied Statistics*, 41(1):142–163.

Laurini, M. P. and Hotta, L. K. (2010). Bayesian extensions to Diebold-Li term structure model. *International Review of Financial Analysis*, 19(5):342–350.

Li, B., DeWetering, E., Lucas, G., Brenner, R., and Shapiro, A. (2001). Merrill Lynch exponential spline model. Technical report, Merrill Lynch working paper.

Ljubojevic, V., Bennett, L.-A., Gill, P. R., Luu, P., Takehara-Nishiuchi, K., and De Rosa, E. (2013). Cholinergic modulation of attention-driven oscillations during feature binding in rats. In *Society for Neuroscience*.

Matérn, B. (2013). *Spatial variation*, volume 36. Springer Science & Business Media.

Matteson, D. S., McLean, M. W., Woodard, D. B., and Henderson, S. G. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, 5(2B):1379–1406.

McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state–space models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, 55(1):199–212.

McCulloch, R. E. and Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88(423):968–978.

Nakajima, J. and West, M. (2013). Bayesian analysis of latent threshold dynamic models. *Journal of Business & Economic Statistics*, 31(2):151–164.

Nason, G. (2016). `wavethresh: Wavelets Statistics and Transforms`. `R` package version 4.6.8.

Neal, R. M. (1999). Regression and classification using Gaussian process priors. *Bayesian Statistics*, 6:475–501.

Neal, R. M. (2003). Slice sampling. *Annals of Statistics*, pages 705–741.

Nelson, C. R. and Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, 60(4):473.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, pages 502–518.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic linear models with R*. Springer.

Piironen, J. and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv preprint arXiv:1610.05559*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.

Polson, N. G. and Scott, J. G. (2012a). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311.

Polson, N. G. and Scott, J. G. (2012b). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

Ramsay, J. O. (2006). *Functional data analysis*. Wiley Online Library.

Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.4.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning. *The MIT Press*.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge University Press.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

Shumway, R. H. and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer New York.

Staicu, A.-M., Crainiceanu, C. M., Reich, D. S., and Ruppert, D. (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics*, 68(2):331–343.

Strawderman, W. E. (1971). Proper bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):385–388.

Svensson, L. E. (1994). Estimating and interpreting forward interest rates: Sweden 1992-1994. Technical report, National Bureau of Economic Research.

Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2):183–204.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

Van der Linde, A. (1995). Splines from a Bayesian point of view. *Test*, 4(1):63–81.

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

Waggoner, D. F. (1997). *Spline methods for extracting interest rate curves from coupon bond prices*, volume 97. Federal Reserve Bank of Atlanta USA.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 364–372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Wand, M. and Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50(2):179–198.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Zhu, B. and Dunson, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *Journal of the American Statistical Association*, 108(504):1445–1456.